New Technologies in Empirical Economics

Inaugural-Dissertation
zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Ludwig-Maximilians-Universität München

2023

vorgelegt von

Valentin Ferdinand Michael Reich

New Technologies in Empirical Economics

Inaugural-Dissertation
zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Ludwig-Maximilians-Universität München

2023

vorgelegt von

Valentin Ferdinand Michael Reich

Referent:	Prof. Dr. Oliver Falck
Korreferent:	Prof. Dr. Andreas Peichl
Promotionsabschlussberatung:	12. Juli 2023

Datum der Promotionsabschlussberatung: 12. Juli 2023
Namen der Berichterstatter: Oliver Falck, Andreas Peichl, Ines Helm

Acknowledgements

During my years as a doctoral student, there were many people that played an important role in completing this dissertation through many means of support. Unfortunately, I can only thank a selection of them here.

First and foremost, I want to thank my supervisor Oliver Falck for his valuable advice, guidance, and the confidence he put in me by offering this opportunity. His comments always helped me to push my work further and he encouraged me to find and consider new solutions. He always had an open door and made time for answering questions. I also thank my co-supervisor Andreas Peichl for his ideas as well as for his inspiring motivation and drive. Furthermore, I would like to thank Ines Helm for completing my dissertation committee.

I am very grateful to my coauthor Anna Kerkhof and my colleague at the EBDC Sebastian Wichert for their mentoring support, advice, ideas, and for always taking their time to listen and to help. There are many things I could learn from them, and discussions with them and their helpful comments made me better understand the process and intricacies of doing economics research. Heike Mittelmeier deserves special thanks for always supporting me throughout my time at the ifo Institute.

I am grateful for my colleagues Lena Abou El-Komboz, Jean-Victor Alipour, Thomas Fackler, Moritz Goldbeck, Simon Krause, and Valentin Lindlacher for helpful comments, inspiration, and productive coffee breaks. I further want to thank many other colleagues and friends from the ifo Institute and the LMU Munich for great discussions, support, and for making these years even more enjoyable.

And most of all, I thank my wife Christina, my son Oskar, and the rest of my family for their love, support, understanding, and especially patience throughout these past years.

Valentin Reich, March 2023

Contents

Preface

New technologies play an increasingly important role in our economy, society, and even in the practice of empirical research in economics. Some of these developments include of course the internet and online platforms. This also enabled the growth in User Generated Content (UGC), i.e., any content created by individuals and published on these platforms. The rise of these platforms and the resulting availability of UGC and other data as a byproduct allowed for new analytical methods to make sense of and profit from these data: *Machine Learning* (ML) refers to models that "learn" patterns from, oftentimes large, data sources. These associations can then be used for example to cluster the data or to make predictions for other data (see Hastie et al., 2009 for an introduction). ML is also frequently applied to texts in the field of *Natural Language Processing* (NLP) to systematically analyze natural language for example to detect its content or sentiment (see Gentzkow et al., 2019 for an overview to applications in economics). Of course, such new methods require specially trained workers to develop and apply them. This led to the emergence for example of the *Data Science* (DS) profession which is about the collection, analysis, and transformation of sometimes unstructured or large data, often by applying ML methods (Kelleher and Tierney, 2018). And these new developments are economically relevant: For example, the World Economic Forum predicted in 2018 that 73 percent of firms will adopt the technology of ML by 2022 such that Data Scientists and related professionals are projected to be among the most requested workers (World Economic Forum, 2018).

These advancements also transpire into the practice of empirical economic research where such techniques are increasingly applied (Currie et al., 2020). Thus, there is the text-as-data literature that uses corpora of natural language, sometimes taken from platforms or other online sources. Other papers use ML techniques for example to classify a variable that is not observed directly. Another technique is *Record Linkage* (RL) which is about linking observations from different micro data sources according to their similarity when there is no common identifier. This process if often aided by ML classification to optimally decide for matching entities. Using these data and techniques can have several advantages: (i) New methods allow to utilize unconventional data sources such as satellite imagery when traditional data sources are unavailable or unreliable. (ii) One can answer questions otherwise not possible due to for example unobserved variables. (iii) In some cases, it is possible to improve data quality for example to find erroneous data or correct for misreporting.[1] (iv) It is possible to get information more timely than with data from statistical offices, sometimes even at higher granularity (e.g., Glaeser et al., 2017).

[1]See Mullainathan and Spiess (2017) and Athey (2019) for an overview over ML and use cases in economics.

At the same time, there are some challenges to using these new technologies for empirical research: First, ML predictions shall be correct on average but individual observations can be incorrectly predicted. Thus, when using ML to predict a variable that is used in regression analyses, there will be some measurement error and it may not be clear to what extent this influences the estimates (Athey and Imbens, 2019). Additionally, the downstream inference can be very sensitive to model choice and this is rarely scrutinized or discussed by practitioners (Ash and Hansen, 2022). Second, both models and data can change over time: The relations learned by models may become outdated as real world behavioral patterns change. This phenomenon is called *concept drift* (e.g., Schlimmer and Granger, 1986) and it might thus pose a risk to external validity. Additionally, data from online platforms can be altered or data access interfaces can be changed such that research may not be replicable in case the data providers prohibit researchers from redistributing the data (Vilhuber, 2020).

This dissertation is about new technologies in both of these senses: On the one hand, it explores how new technologies create new ways for us to interact with our environment, for example to search for jobs in new technology-centric fields or express our views. On the other hand, it uses novel data sources and techniques enabled by technological progress to derive insights.

Many research papers already highlight on the importance of new technologies such as ML and online platforms in the economy and society. One strand of the literature focuses on ML, Artificial Intelligence (AI), and Data Science both in terms of their effects on the labor market but also on potentially negative social effects. Acemoglu et al. (2022) show the relevance of these technologies in the labor market as they find a strong growth in the demand for workers in AI-related professions by analyzing online job vacancies. One potential reason for this development is that these technologies can increase productivity by aiding decision making or through so called *recommender systems*. For example, Kleinberg et al. (2017) show that ML based prediction about criminal reoffending can help judges make better decisions such that the crime rate can be reduced without increasing overall imprisonment. At the same time, AI can also lead to more efficiency by reducing transaction costs as shown by Brynjolfsson et al. (2019). Here, the authors find that machine translation allowed for more international trade on the auction and shopping platform eBay. However, relying on such systems can also bring along societal risks: For example, they can reproduce errors resulting from mismeasurement in the health care system (Mullainathan and Obermeyer, 2017). They can further discriminate women with respect to showing STEM related job ads (Lambrecht and Tucker, 2019) and lead to racial discrimination in criminal prosecution (Arnold et al., 2021). The first chapter of this dissertation contributes to this strand of literature by further studying professionals that develop such technologies. Additionally, I analyze whether public information about their proficiency helps these professionals on the labor market for example because this information can be used by recommendation algorithms in hiring.

Other literature is interested in the emergence of online platforms, such as Farronato and Fradkin (2022), who study the welfare effect of the introduction of the accomodation platform Airbnb on both customers and traditional hotels. The field of labor economics studies work on online labor markets, where tasks such as data entry or data labelling can be commissioned, (e.g., Barach et al., 2020) and other *gig economy* platforms (e.g., Garin et al., 2020). Additionally, some platforms reduce frictions by lowering search

costs. The effects of such a reduction has been analyzed for example in Kroft and Pope (2014): They estimate effects on newspapers after the introduction of the classified ads platform *craigslist* and find fewer classified ads in traditional outlets but no effect on unemployment. Aside from these, online platforms are also studied in various other fields and contexts such as ride sharing (e.g., Liu et al., 2021), advertising (e.g., Decarolis and Rovigatti, 2021), or dating (e.g., Hitsch et al., 2010). My first chapter contributes to this strand of literature by investigating the labor market signaling value of achievements from an online innovation tournament platform.

Another strand of literature revolves around User Generated Content (UGC), i.e., online content that originated from users of a platform rather than being curated content by professionals working for the platform. Oftentimes, this is enabled by social media platforms where users interact with each other and provide each other with content. While this can have some positive effects, for example Fujiwara et al. (2021) show that exposure to the social media platform Twitter led to an increase in voter turnout, there can also be many negative side effects: In particular, there are concerns about addiction (Allcott et al., 2022) and negative influences on mental health (Braghieri et al., 2022). Additionally, it is suggested that social media platforms can increase political polarization, both via bots (Gorodnichenko et al., 2021) and via platforms' algorithms selectively showing content to users (Levy, 2021). Another concern is the exposure of users to hatespeech, discrimination, and bias from UGC. For example, Wu (2018) finds that anonymous posts about women on a discussion forum for economists frequently contain explicit sexual language. Furthermore, Wu (2020) finds gender bias on the same platform where discussions about women revolve more frequently about non-professional characteristics than those about men. Another paper by Müller and Schwarz (2020) finds that an increase in local Twitter usage led to increases in hate crimes in that county. The second chapter of this dissertation contributes to this literature by investigating the extent of gender stereotypes in UGC.

Increasingly, new data and technologies are also used as research tools where scholars analyze unstructured data or data that are a byproduct of other operations rather than research data sets from statistical offices. One such kind are data collected from platforms and websites. This can be either private information that is accessed via a cooperation with the platform or publicly available information that researchers collect themselves.[2] Self collection of public data can happen for example via *webscraping*, i.e., collecting and parsing the information from a website or via an official API, a programming interface to request data. Using such data makes it possible to analyze processes that are otherwise difficult to observe or to get rich data on. For example, Bailey et al. (2018b) use friendship links on the social media platform facebook to compute a new measure for local social connectedness. Such a measure can then also be used to measure the effects of changes in one's social network on decision making in the housing market (Bailey et al., 2018a). Backus et al. (2020) use the wealth of information from transactions on eBay to study different bargaining situations. Another advantage of online platforms is the possibility to run experiments in collaboration with the platform owner. For example, Barach et al. (2020) run an experiment on an online labor market where they varied whether employers could see applicants' wage compensation history or not to estimate the effects on hiring decisions and match quality. Additionally, it is possible to get data in real time such that

[2]Or data from a third party that sells self collected public data.

one can compute economic indicators more timely than official statistics. Thus, Cavallo and Rigobon (2016) collect prices from retailer websites on a daily basis to compute a consumer price index and Glaeser et al. (2017) use data from a review platform for restaurants and shops to measure local economic activity with nearly no time lag.

Another type of unconventional data are corpora of texts which are used in the *text-as-data* literature with the help of various NLP methods. Using textual data allows to measure concepts that are otherwise difficult or impossible to observe. Hence, Giorcelli et al. (2022) are able to observe changes in public discourse, measured by the content and language use in published books, and how it is affected by Charles Darwin's *On the Origin of Species*. Burn et al. (2022) further identify age related stereotypes in job ads and find that these are correlated with hiring discrimination against older men. Another application is to measure how similar objects are: For example, Cagé et al. (2020) cluster newspaper articles to track news events over time and identify plagiarism, ultimately allowing them to measure the returns to original news content. An extensive further overview over the current text-as-data landscape in economics can be found in Ash and Hansen (2022).

While several NLP methods are based on ML algorithms, ML is also applied more generally to other types of data. One use case is to create a variable, be it via prediction or clustering, that is then used in downstream analyses such as regressions. For example, Bandiera et al. (2020) use unsupervised ML to reduce very large and high dimensional data of daily CEO activities to a one dimensional behavior index going from *managerial* to *leadership* behavior. This index is then used as an explanatory variable in a regression to see how it explains firm performance. Furthermore, Record Linkage can be aided by ML classification such as for example in Abowd et al. (2019), allowing to incorporate nonlinearities and more complex patterns for a better linkage. Additionally, ML can even be used as a more general tool for conducting research: For example, Ludwig et al. (2019) show that ML can be used to help credibly extending an analysis under a pre-analysis plan. Another strand of literature in econometrics is *causal machine learning* (see e.g. Athey et al., 2019) which is, however, not the focus of this dissertation.

While the research of all chapters in this dissertation is supported by novel data sources and methods, the use of these techniques and data are not merely alternatives to traditional approaches. Instead, they make these analyses possible where traditional methods would fail. Because the first chapter estimates the causal effect of achievements from a popular online platform for Data Scientists on real world labor market outcomes, the analysis requires data from such a platform to evaluate whether it is valuable for its users. I thus use public data collected from the platform website to analyze dynamics among its users. The focus of the second paper is to measure the prevalence of gender stereotypes in the society since understanding this is an important prerequisite to improve gender equality. Thus, we analyze how the extent of gender stereotypical discussion evolved over ten years by leveraging UGC created under anonymity. This anonymity allows us to overcome the *social desirability bias* (Blackburn, 2017) that would otherwise make the measurement challenging. Hence, we use large amounts of public texts taken from an online discussion forum and measure the occurrence of gender stereotypes in these texts via NLP and ML methods. The final chapter is methodologically focused and describes the linkage of company data from different sources when there is no common identifier. It thus supports the downstream research that uses the resulting linked research

data set. Here, I use traditional data sources but propose that novel methods allow to overcome linkage challenges related to company data. To this end, I explore how ML and NLP techniques can help where simpler methods struggle and potentially make more errors. The following paragraphs introduce these chapters further.

In the first chapter, I analyze the labor market signaling effect of achievements generated on an online platform for knowledge workers. In 2012, Harvard Business Review called *data scientist* the "sexiest job of the 21st century" (Davenport and Patil, 2012). Despite of this, there is hardly a direct path to a data science career as universities only recently started to offer specialized degrees for this field and practitioners try to transition into this interdisciplinary field from various backgrounds. For them, a seemingly adequate alternative channel to demonstrate job relevant skills are online data science competitions where companies seek solutions to their business problems from the crowd. I thus analyze the causal effect of winning achievements, so called *medals*, at a major competition platform for data science. Specifically, I estimate whether these medals increase peer recognition on the competition platform, whether they induce different signaling behavior on participants' resumés, and whether they increase the likelihood of working as a data scientist.

To answer these questions, I use public data collected from the platform combined with the resumés participants made public and linked to their profile. A basic regression of outcomes on winning a medal leads to biased results because winning a medal is endogenous: Higher ability participants are more likely to win and more likely to be successful on the labor market, thus overestimating the effect. I solve this with two distinct identification strategies which allow for a causal interpretation. First, I exploit the sharp discontinuity in the likelihood of winning a medal in a regression discontinuity design (RDD) approach. This is possible because all participants can be ranked on a one dimensional leaderboard per competition and only those in the higher positions end up with a medal. Thus, I can compare the outcomes of those that just won to those that barely did not win. Second, I exploit random variation in the competition organization, where the leaderboard position is dictated by participants' prediction results on evaluation data. This leads to two different scores based on two subsets of this evaluation data. However, only one of them is predetermined to matter for the final placement. Due to random sampling variation in these two subsets, there can be some random variation between the two scores and either score is merely an estimate of the true proficiency. Thus, I use the differences in scores as an instrument for the final leaderboard position in an instrumental variables (IV) approach.

I find that the achievement does have a positive effect on peer recognition within the platform: In the subsequent competition, medal winners are more likely to switch from solo participation to team participation as well as to join a team they never competed with before. Additionally, I find that medal winners are more likely to connect their competition and professional personas by both providing a link to their resumé and mentioning the competitions on said resumé. I interpret this finding that participants do believe the medals have a labor market signaling effect. At the same time, there is some, albeit limited, evidence that participants use the medal to substitute away from other labor market signals. Despite of these findings, there is no significant effect on the likelihood of working as a data scientist. A heterogeneity analysis further reveals that effects are most pronounced for participants coming from backgrounds further away from

data science. For those with a degree in data science, statistics, or computer science, it appears that the medals are even considered to be a negative signal.

Together with Anna Kerkhof, I analyze in the second chapter the prevalence of gender stereotypes. Gender inequality is a pressing issue in our society and the existence of gender stereotypes can be an important influencing factor and obstacle to advancement (Bertrand, 2020). In particular, it has been shown that the existence of such stereotypes can limit both personal and professional development (Jensen and Oster, 2009; La Ferrara et al., 2012; Kearney and Levine, 2015). It is therefore important to understand the extent of stereotypical thinking to find appropriate counter measures. A challenge in this measurement is that individuals may not consciously hold stereotypical views or do not openly express them, knowing that they are not socially desirable (Blackburn, 2017).

To overcome this social desirability bias, we exploit the anonymity of user generated content where individuals can express their views without having to fear real world repercussions. Our data are millions of anonymous public comments to newspaper articles spanning 10 years from a major German newspaper website. NLP and ML methods allow us to measure whether these comments talk about women or men and whether they talk about a gender stereotypical topic. Thus, we can see whether women are more frequently discussed in stereotypical female contexts. We focus on three such contexts: *professional* (stereotypical male), *domestic*, and *physical appearance* (both stereotypical female). To avoid existing gender bias influencing our inference, we develop a new method which combines an unbiased dictionary with word embeddings. This method further allows to make predictions without a labelled training corpus and even works with relatively small dictionaries. It differs from similar recent techniques by allowing for the classification of many different independent topics that are not mutually exclusive.

We find that stereotypes are very prevalent: Women are more frequently mentioned in domestic and physical appearance contexts than men, whereas men are more frequently mentioned in professional contexts than women. There is some decline in gender differences over time for the professional and physical appearance contexts. The gender stereotypes related to household and family, however, are stable over time. These gender differences do not appear to be driven by a difference in news reporting because we find a weaker differential when we apply our method to the newspaper texts the comments are written to. By combining our method with a sentiment analysis, we further find that, on average, women are discussed more positively than men and these differences are largest in discussions with a domestic or physical appearance context. At the same time, however, comments about women are slightly more likely to use offensive language.

The third and final chapter is a methodological contribution describing the linkage of company data sets. Combining entities from different sources into one dataset can greatly expand the utility of each data source. However, linking company data is particularly challenging, for example because firms are often structured in corporate groups. When the linkage serves to create panel data, there are even further complications because companies can restructure, merge, rename, or similar over time. Here, I combine the companies from the *ifo Business Surveys* and the *ifo Investment Survey* with financial data from the commercial *Orbis* database on a micro level. This allows the creation of research datasets containing the rather subjective survey responses alongside companies' objective balance sheet data. The linkage is a major update to one conducted several years prior.

The linkage procedure consists of five steps: (i) *Preprocessing* to clean the data and make it as comparable as possible. (ii) *Indexing* to create a computationally feasible number of possible pairs that shall be evaluated. (iii) *Comparison* to compute similarity metrics for these pairs. (iv) *Classification* to classify pairs as matches or non-matches given their vector of similarity metrics. And (v) *Postprocessing* to make manual corrections and ensure that there is only one match per entity. To address the specific challenges of company data, the steps are tailored to this use case. For example, I use similarity metrics that work well with the properties of company names and explore the use of NLP which is applicable to company linkage. This method results in high match rate, in particular for the most recent years. The match rate is heterogeneous across surveys, partially explained by firm properties which vary by sector such as for example company size. Manual corrections revealed that errors were almost exclusively within-corporate group. They occurred either due to a false link or because a link could not be manually verified when the selection from potential candidates seemed ambiguous.

While the chapters of this dissertation share themes and concepts, all of them are self contained essays that can be read independently.

Chapter 1

Signal or Noise - Signaling Skill among Data Professionals

1.1 Introduction

Degrees in Science, Technology, Engineering, and Math (STEM) can attract students with great career prospectives but skills in these fields become quickly obsolete (Deming and Noray, 2020). At the same time, enabled by advances in information technology, new fields such as *Data Science*[1] emerge in this area, and new professions like these can lack formal education paths. Thus, even though Harvard Business Review called Data Scientist "the sexiest job of the 21st century" (Davenport and Patil, 2012), specialized degree programs are only slowly created such that workers often transition into the field from various backgrounds. And Data Science is economically relevant: It is a fast growing interdisciplinary[2] field which is demanded in all areas of the economy, from marketing, over policing, to health (Kelleher and Tierney, 2018). Additionally, a 2018 World Economic Forum report predicted Data Scientist and related professions to be among the top emerging jobs by 2022 (World Economic Forum, 2018). Thus, if attained college degrees fail to capture current industry trends, they are no sufficient proof for practical skills. Are there other channels for employees to demonstrate their talent to transition into this field?

This study examines online innovation tournaments as one way to signal ability in a context with information asymmetry where traditional education fails to reduce uncertainty about job specific skills. In particular, I study whether achievements in innovation tournaments for Data Scientists serve as a labor market signal. Challenges in such tournaments are often posed by companies because it is a scalable and cost-efficient way to find novel solutions (Boudreau and Lakhani, 2013; Poetz and Schreier, 2012). This means they reflect current industry needs and allow at the same time for a credible and easy to understand signal as they can publicly display how well a prospective employee fared against other specialists.

[1]Data scientists are involved in the collection, analysis, and transformation of oftentimes messy or unstructured data to gain insights, frequently with the help of *Machine Learning* algorithms. In this paper, I refer to occupations that heavily make use of machine learning as *Data Scientist* for simplification. This includes related professions like *machine learning engineers*.

[2]The field is interdisciplinary, as programming skills, statistics knowledge, and domain expertise are required for the profession (Kelleher and Tierney, 2018).

In this empirical analysis, public information of over 250 contests on a major competition platform for predictive analytics is combined with competitors' digital résumés. This allows to estimate changes both within the platform and the real world. Tokens of achievements, so called *medals*, are awarded for high placement in the competitions. These medals are publicly visible and can also be informally referred to in a job search and application process. I thus analyze the effect of obtaining publicly visible medals as a signal for skill. To identify a causal effect, I estimate local average treatment effects by exploiting both the discontinuity in contest placements for obtaining the signal and exogenous variation in contest organization in two separate identification strategies: First, a Regression Discontinuity approach allows to compare the outcomes of close medal winners and losers. Second, I use variation in placement from a random sampling process in the validation of contest submissions as an instrument for medals in an Instrumental Variables approach.

The results show mixed evidence of the effectiveness of the signal. Baseline OLS regressions of the outcomes on a medal winner dummy are intuitive: After winning a medal, individuals are competing more frequently in a team rather than alone. Additionally, winners mention their achievements more frequently on their résumés and they use less non-competition related signals such as recommendation letters. They are further more frequently employed in Data Science positions. However, these results cannot be interpreted causally and the specifications that do allow for causal interpretation can confirm some of these relations: There is a significant effect of reputation from medals on the subsequent likelihood of joining any team or a team with new members. Furthermore, there is evidence that competitors do believe in the effectiveness of the signal as close winners are more likely to link to their professional CVs and more likely to mention the competitions on their CVs. This is more pronounced among individuals who do not already have an academic degree in a related field. However, the effect of of winning on the likelihood of entering Data Science professionally is not statistically significant indicating that participants might compete inefficiently much.

This article contributes to several strands of the literature: It shows another way how platform data can be used for empirical labor economics and contributes to the sparse literature on the relation between online and offline labor markets. More specifically, it is to my best of knowledge the first to analyze how achievements in online competitions translate into real world labor market outcomes. Additionally, it adds to the literature on the sheepskin effect and labor market signals, suggesting that the signaling effect of credentials additional or alternative to formal education may be of some, albeit limited use. Furthermore, it contributes to the literature on the motivation and incentives for open source and crowd contribution by showing that participants of online competitions do at least in part believe they can benefit from this signal.

The outline of this paper is as follows: Section 1.2 gives an overview over preceding research. To better understand the empirical setting, the platform is described in section 1.3, the data in section 1.4, and the estimation strategy in section 1.5. Estimation results are presented in section 1.6 and discussed in section 1.7. Section 1.8 concludes.

1.2 Related literature: From offline to online signaling

In terms of standard education signaling, one can best compare this paper to results from the literature on the *sheepskin effect*: the signaling value of holding a degree, irrespective of the years spent in education and accumulated human capital (Jaeger and Page, 1996). An article on the sheepskin effect related to this study is Clark and Martorell (2014). Using a Regression Discontinuity Design, the authors show that for students with similar scores, those that barely pass a high school exit exam fare better at the labor market than those that barely fail. Other empirical studies can confirm the relevance of a degree certificate, indicating that employers might care about a simple and easy to screen signal (see Caplan, 2018).

It is long acknowledged that, even among offline options, traditional education is not the only way to obtain a labor market signal. Some literature looks at the general equivalency diploma (GED), an alternative for students that dropped out of high school, and finds that holders are better off than without but it is still a weaker signal than an actual High School diploma (see Stanley et al., 1998). Post-baccalaureate business certificates, aimed at students with a non-business major, additionally capture the aspect of interdisciplinarity. However, these could not be easily associated with positive effects in hiring (Gaulke et al., 2019).

Research is also interested in online generated digital signals such as involvement in Open Source Software (OSS).[3] One motivation to contribute to community projects with unpaid work, are career prospects (Lerner and Tirole, 2001; Leppämäki and Mustonen, 2009; Bitzer et al., 2017) but not necessarily the key driving factor (Lerner and Tirole, 2005; Athey and Ellison, 2014). Orman (2008) finds that OSS involvement can act as a signal to receive higher wages when combined with a college degree but cannot replace traditional education. Conversely, Hann et al. (2013) find that OSS contribution is an effective and credible signal that can be related to substantial wage increases. For a contribution to be an effective labor market signal, it and its relevance must be well visible outside of the OSS community but the credibility also depends on the size of the project and the how well the refereeing process functions (Lee et al., 2003). And while there are mechanisms in place to track each individuals' contribution and its relevance (Lerner and Tirole, 2005), this may not be as straightforward to see for a hiring manager as it is for peers. Two studies very related to mine are Xu et al. (2020) and Huang and Zhang (2016). Both make the connection between reputation from online crowd contribution and real world labor market events. The first combines data from a large software question-and-answer platform and a job board platform. The findings suggest that contribution is to some degree driven by career concerns. The second further shows that contributing to a specific Open Knowledge community affects the likelihood of job changes by combining data from the crowd sourced community and résumés from the professional social network LinkedIn.

Another strand of the online signaling literature revolves around Online Labor Markets (OLM), two-sided market platforms such as *Mechanical Turk* and *Upwork*, to offer and find freelance work. Empirical studies using data from such platforms find that users

[3]See Osterloh and Rota (2007) for a history of OSS and a survey of the OSS literature.

can particularly benefit from increased transparency and standardization of information. Pallais (2014), for example, shows that more publicly available info about themselves and their work history helps relatively inexperienced workers to be hired. Similarly, Stanton and Thomas (2016) find that agencies can provide workers with a signal that helps finding a first job on the platform. Agrawal et al. (2016) estimate that the employment probability is higher when there is standardized and verifiable info on employees, especially for otherwise disadvantaged groups from less developed countries. Skill certificates can function as a signal for workers without long working histories according to Kässi and Lehdonvirta (2019). Other studies can confirm the positive effect of signaling within the platform (e.g. Barach et al., 2020; Horton and Barach, 2020). All of these articles estimate the effect of signaling on success within the platform and do not identify potential impacts on outside options. Interestingly, Claussen et al. (2018) show that traditional offline signals can become meaningless in OLMs relative to online generated ones.

The literature about innovation contest platforms focuses mostly on the dynamics and determinants of competitor effort and success (e.g., Garcia Martinez, 2017; Wooten, 2022; Dissanayake et al., 2018; Dissanayake et al., 2019; Lemus and Marshall, 2021) and on competition design to optimize performance (e.g., Boudreau et al., 2011; Bockstedt et al., 2016; Wooten and Ulrich, 2017). Archak (2010) shows that achievements are an effective signal within the platform, in the sense that reputation is used strategically by high ranked users to deter others from competing. Other than that, I am not aware of studies that further explore the signaling effect in innovation contests.

Overall, the literature indicates that there is potential for a labor market signal from online innovation platforms but there is no conclusive evidence on its effectiveness yet. Alternative offline signals appear to be effective to some degree albeit less than their traditional counterparts. The comparable involvement in OSS projects to demonstrate skill does appear to be beneficial, and standardized info on OLMs is effective but nothing can be said about its effectiveness outside of the platform. Hence, I am interested in whether an easily verifiable and standardized signal generated on an online platform has a positive effect on offline labor market success. In the next section, the respective platform and the signal shall be described in detail.

1.3 Background: Contest description

For this study, I collected data from a major data science innovation contest platform. Over several years, the website attracted millions of users overall and more than 120.000 users[4] to compete in over 350 public data science competitions. Among the organizations that use it to search for innovative solutions from the crowd are large and well known firms for example from technology, financial, and manufacturing sectors.

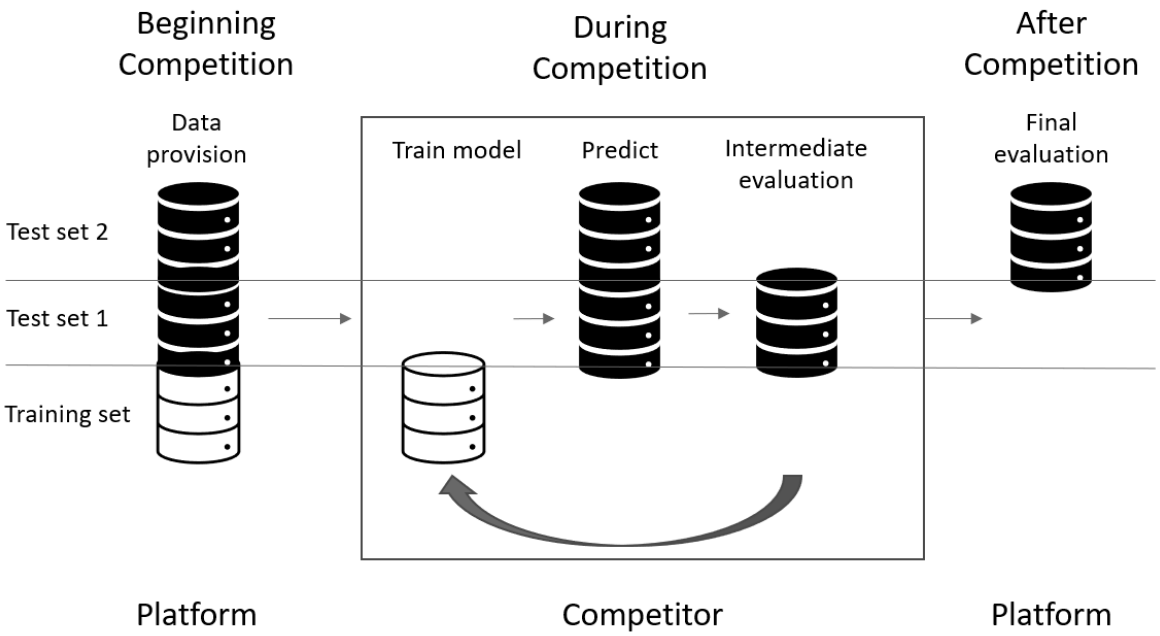
The website is an ideal platform to study this research question for five reasons: (i) it is one of the largest and most well known competition platforms in Data Science, a field where I expect additional signals to be particularly valuable due to its interdisciplinarity. (ii) In 10 years, only around 300 competitions handed out medals, making the signal relatively scarce. (iii) Since everything is public and all submissions can be ranked on

[4]The number of overall users is larger because many registered users only use the forum or download datasets.

a one dimensional leaderboard, good placements and achievements are relatively easy to acknowledge as well as to see by and communicate to recruiters, HR personnel, or hiring managers with less technical involvement in the field. (iv) The average competition lasts 80 days and usually requires continuous effort by the participants, thus displaying the job relevant ability of competitors to work, sometimes in a team, on a larger project. And (v), many challenges are posed by actual companies, meaning that the tasks are industry relevant.

Most competitions follow the same workflow and rules: The ultimate goal is to solve a prediction task with provided data. Tasks can range from classifying weather and vegetative properties in pictures of forests to identifying toxic comments on online forums. Whoever makes the best predictions on a specific data set wins the competition. One can participate alone or in a team of up to five people.

Figure 1.1: The competition workflow



Note: A typical competition workflow. The white training data set has labels (i.e., the variable to predict) known to both competitors and organizers ("Platform"). The two black test data subsets are both included in the same test data file and only the organizers know their label. The competitors are not able to tell the two test data subsets apart.

A typical competition workflow with three stages is visualized in figure 1.1. In the first stage, the contest organizers create multiple files that the competitors can download and use for the competition. The files typically represent two types of data sets: training data and test data[5]. Both contain a matrix with the same set of covariates (the *features*)

[5]Depending on the competition details and the medium of the instances — tabular data, text, image, sound files, ..., or a combination of these — the training and test data sets can each consist of multiple files and there can also be supplementary files. For simplicity, I refer only to one training data and one test data file and assume tabular data as well as a classification task for my explanations. However, the concepts hold just the same for other types of data and regression tasks.

and they differ in whether they additionally contain the variable(s) to predict (the *labels*): The training data contain labels visible to both competitors and organizers whereas the labels of the test data are known only to the organizers. Additionally, the organizers separate the test data into two subsets, call them *Test set 1* and *Test set 2*. Competitors only know that the observations are somehow subdivided into these two sets but they cannot distinguish them.

In the second stage, during the competition, competitors use any method to make predictions for all of the test data, usually by training machine learning algorithms. At any point during the competition, they can upload a file containing the IDs of all of the test data and their predicted labels. They will then be evaluated automatically by how close they are to the true labels[6] but only on a subset of the test data, namely *Test set 1*. The competition does not end here. Competitors can see an intermediate score and their position on an intermediate leaderboard from their submission as feedback how well they perform relative to other teams. Then they can improve their model and repeat this stage up until the competition deadline.

In the final stage, after the competition deadline, teams are automatically evaluated on the other subset, *Test set 2*. Only this evaluation matters for determining the final winner.

Since every submission is evaluated with a single scalar metric, it is straightforward to rank all participants on a publicly visible leaderboard in real time. The position on the intermediate leaderboard is given by the score on *Test set 1*[7]. It only serves as an approximate indicator for how well one performs. The final leaderboard is given by the evaluation on *Test set 2*.[8] Switching the relevant data subsets from intermediate to final evaluations ensures that teams can experiment to improve their algorithm with feedback but they cannot game the system[9]. As a consequence, a substantial shuffle in the leaderboard can occur at the end, and participants may suddenly and unexpectedly climb the ladder to the top or fall towards the bottom.

The first couple of winning teams usually win a monetary prize but even if one does not expect to be at the very top, there is an incentive to compete: The leaderboard position shows how well one performs relative to others. This makes it a credible and easy to understand signal of relative skill. Furthermore, so called *medals* are awarded for high placement on the leaderboards. The best x%[10] of teams in any competition receive such a digital token as a reward[11] which is displayed on the website's user profile. Because these publicly visible medals enable a quick way of verifying someones relative data science and

[6]For the evaluation, a scalar metric will be computed given by some function of true labels and predicted labels. This metric varies from competition to competition and can be for example the accuracy, the mean squared error etc.

[7]Competitors choose with each submission the one that shall count, i.e., the one that will be used for both final and intermediate evaluation in case no further submissions are made.

[8]Both leaderboards are publicly visible once the competition ends.

[9]If the intermediate and final evaluation were based on the same observations of the test data, one could get perfect scores by iteratively varying the target values starting from a random guess and checking the score. This way, one would eventually find out for each observations what its true value must be. In reality, this would be impractical because the number of daily submissions is limited and because this strategy will most likely make the participant worse off in the final ranking.

[10]This share is not fixed and depends mostly on the size of the competition. Table A1 in the appendix shows how many medals are awarded for differently sized competitions.

[11]Depending on their placement, competitors receive a bronze, silver, or gold medal.

machine learning skill, I argue that users can use them informally as a signal on the labor market for data scientists by sharing their profile during the application process.

The focus of this study is on the effect of a bronze medal rather than silver or gold. This is done to estimate the effect of obtaining *any* signal as opposed to none, while for silver and gold this is not straightforward to do. For some of the identification strategies, these can only be used to tell us the effect of an *incrementally better* signal. In section 1.6.5, this will be further discussed.

1.4 Data

Competition platform data Most of the data for this study have been collected from the platform website. At the time of collection, 287 public medal awarding[12] competitions have been finished on the platform. Table A2 in the appendix describes these competitions. The median competition has just under 400 participating individuals in 351 teams. Most competitions go on for two to three months and during this time users can submit up to 5 result files per day. On average, 172 competitors receive a medal per competition.

Labor market data Competitors' real world labor market outcomes come from their profiles on the professional social network LinkedIn. Many individuals added a hyperlink on their competition website profile to allow visitors to find their page on the professional network. One can think of a LinkedIn profile as a digital résumé which is openly visible in a network of more than 600 million international users (LinkedIn, 2020). The network can for example be used by recruiters to find talent or by employers screening job applicants. Aside from education and past work experience, jobseekers can list various other relevant elements such as skills, accomplishments, and interests. Thus, the data contain info not only on competitors' professional history but also on other signals they use. Profile data from LinkedIn have previously been used for empirical research for example by Huang and Zhang (2016) and Xu et al. (2020).

The data only represent information that is publicly available and need to be interpreted accordingly. Other profile information may only be visible to direct connections. Hence, when I do not observe, for example, any academic degrees, I cannot distinguish whether this is because the individual has not listed any or because she has not made the info public. Likewise, a lack of skills listed on LinkedIn of course does not imply that the person has no skills in the real world. For this reason, the *number of skills* variable, for example, does not proxy for how skilled a person is but for how many skills she deliberately signals to the public. It is thus a measure for *signaling* skill rather than *actual* skill. While profiles from professional social networks do not necessarily show the entire truth of the working history, they indicate what workers want to show prospective employers. This allows to reduce knowledge asymmetries and users have an incentive to fill their info as good as possible with carefully selected signals.

Descriptives The final data set is a panel with the publicly available competition leaderboard rankings. This means an observation is an individual at a specific ranking

[12] Additional non-awarding competitions may either serve as easier testing and learning grounds or they feature fun or very peculiar challenges posed by the website organizers with simulated data.

Table 1.1: Descriptive statistics

	Full sample			Linked sample		
	Observations	Mean	SD	Observations	Mean	SD
Final position	253112	0.45	0.28	22749	0.30	0.24
Score change	233058	-0.01	0.10	20652	-0.01	0.09
Medal winner	253112	0.18	0.38	22749	0.32	0.47
Bronze	253112	0.08	0.27	22749	0.15	0.35
Silver	253112	0.08	0.27	22749	0.13	0.34
Gold	253112	0.02	0.14	22749	0.04	0.19
Money	253112	0.01	0.08	22749	0.01	0.11
Team size	253112	1.41	0.92	22749	1.39	0.91
Experience (competitions)	253112	4.14	10.49	22749	11.72	17.87
Submissions	249504	19.70	38.37	22501	32.44	55.55
Used public code	253112	0.02	0.16	22749	0.03	0.18
Published own code	253112	0.01	0.10	22749	0.01	0.10
Outcomes:						
1) Team formation:						
Competes in team	253112	0.21	0.41	22749	0.20	0.40
New team	253112	0.19	0.39	22749	0.16	0.37
Switch to team	89783	0.10	0.30	13961	0.14	0.35
2) Signaling:						
Links LinkedIn	160406	0.30	0.46	22749	1.00	0.01
Mention competitions	22753	0.72	0.45	22749	0.72	0.45
Certificates	22753	2.74	8.04	22749	2.74	8.04
Recommendations	22753	1.08	3.08	22749	1.08	3.08
Skills	22753	5.99	12.51	22749	6.00	12.51
3) Labor market:						
Data Scientist	22753	0.54	0.50	22749	0.54	0.50

Note: *Final position* is the relative position on the final leaderboard and ranges from 0 (best) to 1 (worst). The mean is below zero because the position varies on a team level and team sizes are larger on the better ranks. *Score change* measures the change in prediction scores from intermediate to final evaluation. *Money* is a dummy indicating the top participants who win a monetary prize. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession.

position in a competition. It is possible to track competitors over time in this unbalanced panel if they compete more than once. For a sample of competitors, info from their LinkedIn profiles has been manually collected. This sample is selective in the sense that it only includes competitors that would contribute to the identification of the local average treatment effects described in the next section.[13]

Table 1.1 shows descriptive statistics. The left set of columns reflects the full sample of users, while the right columns show statistics for users that could be linked across the two data sources. The panel contains more than 250.000 observations from around 120.000 individuals in 287 competitions. Most entries stem from solo participation and 59 percent come from individuals that never compete in a team with others. Teams are capped at a size of 5 with 1.4 members[14] on average. Figure A1 in the appendix further shows how the 21% of observations that participate in a team with other participants are distributed across different team sizes. In the linked sample, we can see that competitors here are effectively the better and more experienced ones. We can also see that 72% of users mention the competition website anywhere on their LinkedIn profile[15], hinting towards some role of competitions in the job search process. Around 54 percent work as data scientists[16]. Several profiles do not include any info about skills — self reported and sometimes endorsed by others — but those that do, list various technical and non-technical skills ranging from specific programming languages to public speaking. Furthermore, figure A2 in the appendix shows that one third of the participants competes more than once such that they compete 2.1 times on average.

1.5 Empirical strategy

1.5.1 Identification

Baseline A basic way to estimate how an outcome in period $\tau > 0$ relates to winning a medal in competition c_0 in period $\tau = 0$ for individual i is via the baseline equation 1.1.

$$y_{i\tau} = \theta^{ols} Medal_{ic_0} + \beta_1^{ols} X_{i\tau} + \beta_2^{ols} X_{ic_0} + \gamma_i^{ols} + \gamma_\tau^{ols} + \gamma_{c_0}^{ols} + \epsilon_{i\tau} \tag{1.1}$$

$$Medal_{ic_0} = \mathbb{1}[i \text{ has won a medal in } c_0] \tag{1.2}$$

$Medal_{ic_0}$ is a binary indicator of value 1 if the individual won a medal in c_0 and 0 otherwise. Observed factors that vary over both time and competitors are controlled for via the vectors $X_{i\tau}$ and X_{ic_0}. The parameters γ_i, γ_τ, and γ_{c_0} capture individual, outcome period, and treatment competition fixed effects respectively. The coefficient of interest, θ^{ols}, thus measures the within competitor change in the outcome after winning.

[13]For identification with the Regression Discontinuity approach, all observations within a specific threshold have been sampled. For identification with the Instrumental Variables approach, a random sample has been drawn. Both of these steps were taken to avoid collection of unnecessary data.

[14]Here, the term *team* refers to individual leaderboard positions and includes "teams" of one.

[15]Mentions can range from linking to their profile, referring to themselves as competitors on the site, to describing the taks and their results in the *Accomplishments* section. Some even list their participation in competitions as work experience.

[16]This includes related roles in the predictive analytics domain such as data engineers, machine learning engineers, etc.

The results can be biased if there are unobserved factors that vary over both time and competitors, such as motivation. However, the way competitions are organized introduces two sources of exogenous variation that allow for causal identification by estimating different local average treatment effects. A comparison of the two identification strategies and their respective advantages can be found in appendix section A.2.1.

Winning discontinuity First, I exploit the sharp cutoff between those on the leaderboard that do and do not receive a medal using a sharp Regression Discontinuity Design (RDD). For example, the top 10% of teams on the competition specific leaderboard win, and everyone with a lower rank loses. Assuming that competitors in a close window around this threshold are approximately equally talented, differences in their subsequent outcomes arise only due to whether they won a medal or not. The distance from the minimum rank to obtain a medal, relative to the size of the leaderboard, serves as the running variable. One can then interpret this Local Average Treatment Effect (LATE) as the effect of just closely winning a medal where participants are likely very homogeneous in terms of their skill level.

The assumptions for a valid RDD are that individuals to the left and right of the threshold are comparable and that they cannot control perfectly what side they land on. Table A4 in the appendix shows that competitors on both sides of the cutoff were not systematically different from one another. Neither based on various attributes in their preceding competitions (left column) nor on attributes at the time of the competition (right column). Additionally, perfect control is impossible: For one, because the final position depends on both a changing numbers of participants and the unpredictable rank change, but also because that would require control over the other competitors as well. Furthermore, following McCrary (2008), the distribution of the running variable in figure A4 shows that there is no meaningful discontinuity or bunching at the cutoff.

Here, a first order polynomial line is fitted on both sides of the medal winning score cutoff. Equation 1.3 shows the how the RDD is estimated with the coefficient of interest being θ^{rdd}, the jump in fitted lines at the threshold:

$$\begin{aligned} y_{i\tau} = \theta^{rdd}\text{Medal}_{ic_0} + \beta_1^{rdd}\text{Distance}_{ic_0} + \beta_2^{rdd}\text{Distance}_{ic_0} \times \text{Medal}_{ic_0} \\ + \beta_3^{rdd}X_{i\tau} + \beta_4^{rdd}X_{ic_0} + \gamma_\tau^{rdd} + \gamma_{c_0}^{rdd} + \epsilon_{i\tau} \end{aligned} \tag{1.3}$$

The further one moves along the running variable *Distance* away from the threshold, the more dissimilar the competitors appear. Hence, a triangular kernel weighting function gives lower weight to observations the further they are away from the threshold and only a range of 2.5% of the leaderboard on both sides respectively is considered.

Score change instrument The second identification strategy uses an instrumental variable (IV) approach. A valid instrument must be correlated with the likelihood of earning a medal but not be related to subsequent outcomes in other ways. The change in the base data to compute the evaluation metric at the end of each competition satisfies these conditions. As we have seen in Section 1.3, users are evaluated on two samples of the test data. Thus, for the final submission, there are two different scores, the *intermediate* and the *final* score.[17] of which only the latter matters. The difference in these scores

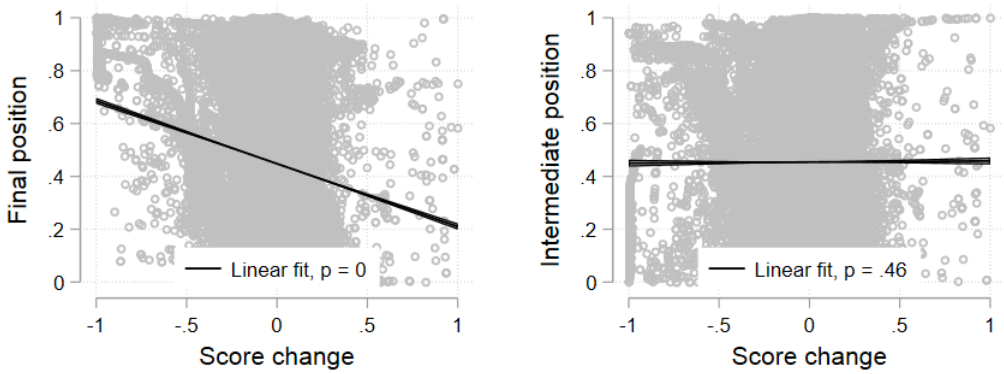
[17]Even though both scores are determined from the same final submission, I call them *final* and *intermediate* because the intermediate score is displayed already before the competition deadline.

is correlated with the likelihood of obtaining a medal, i.e., is a relevant instrument: If a competitor's score sufficiently improves with the change in evaluation data, she moves up on the leaderboard and is more likely to win. Figure A3 shows that participants in the higher ranks frequently do or do not win a medal due to the score change. Here, the LATE can be interpreted as the effect of a medal won by chance where the probability of being a complier is distributed with a bell shape around the RDD cutoff. Hence, there is substantial overlap in the observations that contribute to the two LATEs but the IV is additionally identified from a few individuals that end up further away on the leaderboard.

The instrument is exogenous under the assumption that the two subsets of the test data are both samples drawn at random from the same population and thus not correlated with the other variables. Evaluation of neither subset represents the true quality of the trained algorithm and thus competitor skill but rather an estimate thereof. Thus, in expectation, the score on both subsets should equal the true score, i.e., the score one would obtain on the population the samples are drawn from: $E[score_{true}] = E[score_{test1}] = E[score_{test2}]$. Hence, I argue that any difference in scores is only due to random sampling variation in the two subsets.

The scatter plots of figure 1.2 confirm this: The plot on the left shows that score changes are significantly related to the position on the final leaderboard, thus supporting the relevance assumption. At the same time, the plot on the right shows that score changes are not related to the intermediate leaderboard position, i.e., the one resulting from the other evaluation data subset, supporting the exogeneity assumption.[18]

Figure 1.2: Score change against position on the leaderboard

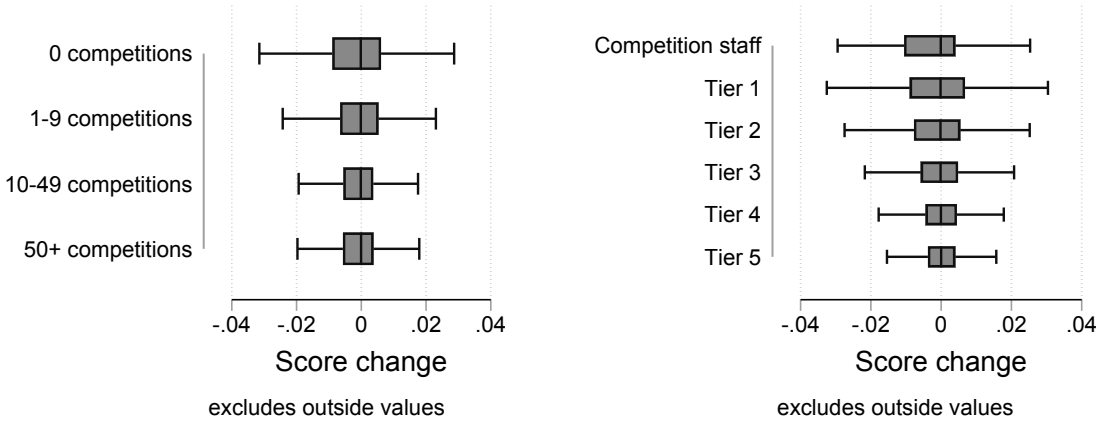


Note: Scatterplots with changes in min-max transformed scores on the x-axis and the leaderboard position in percentiles on the y axis. Left image: position on the final leaderboard. Right image: position on the intermediate leaderboard. The dark line shows the linear fit and its 95% confidence band. Standard errors clustered on team-competition level.

The score change is related to skill only in the sense that more skilled competitors' algorithms generalize better to different data sets and thus *absolute* score changes are smaller. Less experienced competitors should have higher variance in their score change

[18]Because the intermediate score is one estimate for skill, this implies that the score change is not correlated with skill.

Figure 1.3: Distribution of score changes by levels of experience



Note: Boxplots showing the distribution of scores by different levels of experience. Left image: experience in terms of competitions. Right image: experience tiers. The "tier" is a coarse ranking of users' proficiency on the website, determined by the number of medals won.

but importantly: the sign of the difference is ambiguous[19]. Figure 1.3 shows how the score change is distributed among competitors with different levels of competition experience on the left and different skill groups, determined by the number of medals, on the right. As hypothesized above, they have approximately the same average score change of zero but the less experienced groups have a higher variance. Scores are normalized to be in the [0,1] range within each test set to look at relative score changes[20].

Equations 1.4 and 1.5 respectively show the second and first stage of the Two Stage Least Squares estimation:

$$y_{i\tau} = \theta^{iv}\widehat{Medal}_{ic_0} + \beta_1^{iv2}X_{i\tau} + \beta_1^{iv2}X_{ic_0} + \gamma_\tau^{iv2} + \gamma_{c_0}^{iv2} + \epsilon_{i\tau} \tag{1.4}$$

$$Medal_{ic_0} = \eta^{iv1}ScoreChange_{ic_0} + \beta_1^{iv1}X_{i\tau} + \beta_1^{iv1}X_{ic_0} + \gamma_\tau^{iv1} + \gamma_{c_0}^{iv1} + \epsilon_{ic_0} \tag{1.5}$$

The first stage regresses the indicator for a medal on $ScoreChange$ - the change in evaluation metrics from the intermediate to the final test data - and a set of covariates and fixed effects. The second stage is largely equivalent to equation 1.1, except that it replaces the binary $Medal$ variable with its prediction from the first stage. The coefficient of interest is θ^{iv}.

[19]It is, however, possible for a participant to systematically overfit to the intermediate leaderboard, i.e.i, one split of the final submission file. Then, the difference is not random anymore. There is only one possibility to overfit to the split of the intermediate evaluation. Only when participants repeatedly submit, check the score, and adjust their model accordingly, can they learn about this subset of the data. This is not only a strategy that would lower chances of winning and should thus not be expected from rational participants, it is also something one can account for by controlling for the number of submissions.

[20]Scores are min-max scaled using the formula $\frac{x_i - min(x)}{max(x) - min(x)}$. This helps making the scores comparable across leaderboards and competitions.

1.5.2 Outcomes

1.5.2.1 Competition outcomes

Outcomes from the competition platform are used for the first research question, i.e., whether competition achievements build reputation on the website itself. The hypothesis is that winners have better opportunities to join teams with others in subsequent competitions because the medal helps them convince their peers of their proficiency. Other competitors may be more reluctant to form a team with someone when there is uncertainty about her skill. Teams can form and merge during the competition and while it is of course possible that team members know each other and organize in the real world, every competition has its own discussion board to talk about different aspects of the challenge. Here, competitors can get to know each other and often there is one or more well frequented threads specifically dedicated to finding team members.

The first outcome, *competes in team*, is a simple dummy indicating whether the user participates in a team rather than solo.[21] The second outcome, *new team*, is a dummy indicating that an individual competes with a team where they have never competed with any of the other team members before on the platform.[22] The third outcome, *switch to team*, is a dummy indicating that an individual competed solo in the previous competition c_0 and competes in a team with others in c_τ.

1.5.2.2 Resumé outcomes

To analyze the effects of obtaining the signal on outcomes that go beyond the competition platform, i.e., on signaling activity and labor market success, the data requires to use outcomes that do not vary over time: While the LinkedIn profiles do contain users' employment and education history, a lack of standardization makes these difficult to use. For this reason, I only use information about current employment which appears to be of higher quality[23]. Likewise, I can only observe the signals that individuals use on their digital résumé at the time of data collection. This information is assumed to represent the state of the competitors *after* the competitions since it has been collected after the competition data. Hence, because there is no time variation within individual, it is not possible to use panel methods such as within estimation, and all individuals in the data are treated as independent observations in a repeated cross section.

The first question to answer with these outcomes is to find out whether résumés of medal winners differ significantly in the number and type of signals from non-winners. The hypothesis is that winners are more likely to use the achievement as a signal. This is measured by whether they provide a hyperlink on their competition website account to LinkedIn and by whether the name of the competition website can be found in their

[21] Thus, this variable is 0 if they participate solo and 1 if they participate together with other competitors.

[22] This variable is undefined when they participate solo. It takes the value 1 when they participate with others they never competed with before and 0 when they were in a team with any of the other team members before.

[23] For example, in the history, dates can be missing, participation in competitions is mentioned as work experience, and education can contain both university degrees and online course certificates. At the same time, it is not clear how long it would take for a signal to be effective and using the most recent information should be more informative than looking at immediate effect.

profile page on LinkedIn.[24] Additionally, I estimate whether medals are a substitute or complement to comparable signals by using the number of certificates, letters of recommendation, and listed skills as outcomes. These are comparable in the sense that they are not expected to replace traditional education but rather give additional information about job competence. When acquiring any of these signals takes time and their values are independent from one another, the hypothesis is that these competition-unrelated signals are substituted away from.

The other question to answer using CV outcomes is whether medals are effective and help competitors enter the field of Data Science. For this, I use a dummy indicating whether the person is currently working as a data scientist. Here, changed behavior with respect to unrelated signals could lead to an omitted variable bias. If medal winners substitute away from other non-competition related signals, the total effect on labor market success is biased towards zero. To address this, I include the non-competition related signals used as outcomes for the second research question as controls here. If I did not account for this, I would not capture any signaling effect if, for example, the participants were to use the signaling value of medals only to reduce their number of letters of recommendation while keeping their level of employability.[25]

The effects of medals on actual employment are Intention To Treat (ITT) effects because the model does not incorporate information about whether the individual uses the signal. This could be imperfectly achieved by using the previously mentioned measures for signal usage. However, it is not observable what kind of information competitors mention informally on their actual job application documents outside of LinkedIn. For this reason, I abstain from interacting the treatment with a signaling dummy.

A complication with the time invariant outcomes is that there are circa 300 independent treatments, one for each medal awarding competition, but there is only one outcome per individual. In each of these treatments, a given participant can win a medal and thus be in the treatment group, the control group, or neither.[26] This means that any participant can be treated multiple times or even contained in both a treatment *and* control group of different competitions. If competitors in the control group of one competition are in the treatment group of another and the outcome is measured only after all of the competitions, then the treatment effect can be biased: I would compare the outcome of a treated individual to that of an individual who is also treated by the time the outcome is recorded. Table A3 shows that over all competitions, around 1400 (1600) competitors of the full sample have been in both groups of the RDD (IV) at least once. To deal with this challenge, these individuals are excluded in the estimation.

1.6 Results

This section presents the results from the estimations. First for the effectiveness of the signal on the formation of teams within the platform and then for the external use and

[24]Either of these measures show that the participant makes a connection between their competition and professional personas. I.e., they want vistors from the competition website to consider them in a professional context and visitors of their résumés to view their competition achievements.

[25]I also present results without these controls in this paper.

[26]For the IV, the analog is to interpret the "treatment group" as the compliers that won due to a score change and the "control group" as the individual that did not win due to a score change.

effectiveness of the signal.

1.6.1 Effect on team formation

Table 1.2: Estimation results: Effect on team formation

	(1) Competes in team	(2) New team	(3) Switch to team
a) OLS:			
Medal winner (t-1)	0.01619***	0.00865**	0.01446***
	(0.00398)	(0.00361)	(0.00384)
Observations	93468	93468	78153
Individuals	29296	29296	25381
b) RDD:			
Medal winner (t-1)	0.04228**	0.03482**	0.03421*
	(0.01980)	(0.01690)	(0.01871)
Observations	7648	7648	6163
Individuals	3768	3768	2978
c) IV:			
Medal winner (t-1)	0.05135**	0.00329	0.02154
	(0.02372)	(0.02135)	(0.02375)
Observations	100602	100602	85181
Individuals	28393	28393	24671
1st stage F	1620.09	1620.09	1422.27

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

I start by reporting results for the baseline OLS estimation, i.e., a regression of team participation indicators on a medal dummy, in Table 1.2, panel a). The table displays how success in one competition is related to outcomes in someone's next competition. The first column contains results for a *team participation* dummy outcome and the second for the indicator of being in a *new team*, i.e., with new team members. The third column contains results for a *switch to team*, i.e., whether participants competed solo before and now in a team with others. Coefficients from the linear probability model can be interpreted as percentage point increases in the likelihood of the outcome occurring after winning a

medal. Column (1) shows that winning a medal significantly relates to a 1.6 percentage point higher likelihood of subsequent team participation. As table 1.1 showed, on average, around 20% of observations is team participation. While the coefficient for *New team* in column 2 is approximately half as large but still significantly positive, column 3 shows that winning a medal is related to a 1.4 percentage point higher likelihood of switching from solo to team participation. As discussed before, this specification does not allow for a causal interpretation and hence shall be re-estimated using the other identification strategies.

Results from the comparison of close winners and losers via an RDD are reported in table 1.2 panel b) and they represent the jump of the fitted lines at the medal winning threshold. The coefficients are substantially larger than the baseline results, due to the inclusion of individual fixed effects in the baseline.[27] The effects are significant on the 5% level for any team and new team participation and on the 10% level for a switch from solo to team participation. Figure 1.4 also shows the RDD plots for the outcomes with clearly visible and sizeable discontinuities with increases of more than 30%. For all outcomes, the slopes of the fitted lines switch signs after the cutoff. The positive slope to the right is natural due to the positive correlation between leaderboard position and the likelihood of competing in a team. A possible explanation for the negative slope to the left is that not winning so close to the threshold could even have a negative signaling effect: Looking at the leaderboard on the website makes the loss visually more obvious when the cutoff is displayed on the screen above the participant at the same time. This may have a psychological effect on potential team members.[28]

An alternative way to estimate a Local Average Treatment Effect arises from using the score change as an instrument. Its results confirm the baseline results at least partially as table 1.2 shows in panel c). Winning a medal leads to a weakly significant increase of team participation of around 5 percentage points which is around one quarter of the sample mean. The results for *new team* and *switch to team* are not significant. More details about the first stage can also be found in a dedicated regression table for the IV in the appendix table A5.

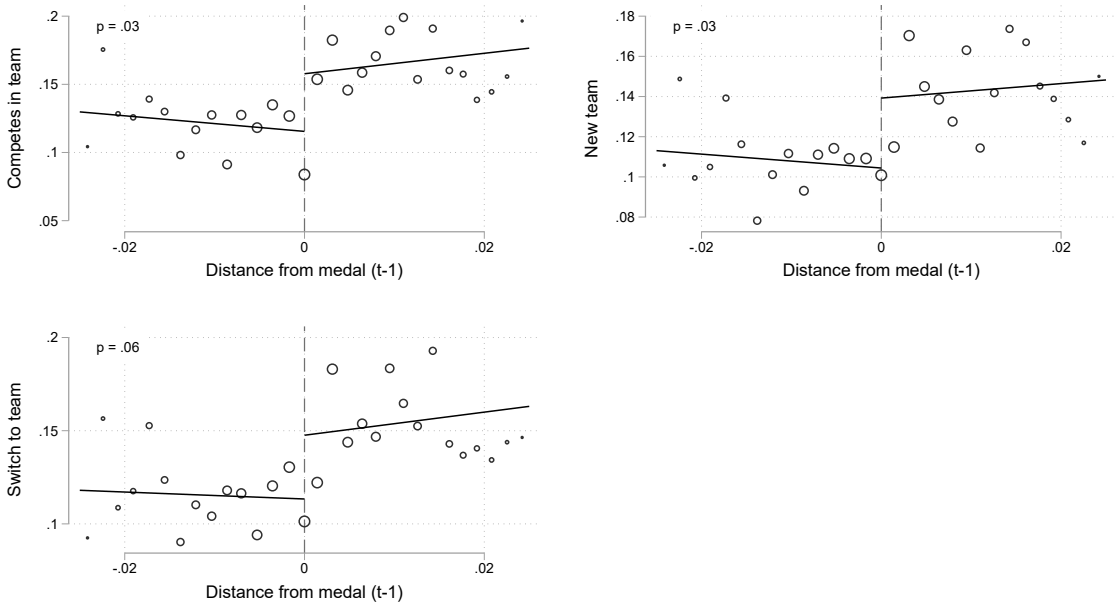
It is also possible that it takes more time for the signal to come into full effect, such that it is too shortsighted to only look at outcomes in someone's next competition. For this reason, I separately estimate the effect of winning a medal on outcomes of the next few following competitions. Figure A5 shows the coefficients for outcomes at different points in time after competition c_0 for the baseline estimation, the RDD, and the IV. Almost all specifications show a significant immediate effect that lasts only for the next competition. The only exception is the effect on joining a new team estimated with the IV strategy in panel (f), where there is no clear effect.

In sum, there is some evidence that achievements on the platform cause immediate team building benefits. If the signal works among peers, it is possible that participants

[27]With individual fixed effects, the effects boil down to changes in differences from the individual mean. However, if medals lead to persistent effects on team formation, the outcome variable of future observations, and thus their mean, will be affected as well. Thus, the effect, as a difference from the mean, is lower in absolute terms. Figure A5 confirms this possibility by showing that, even if not significant for each individual one, the team formation effects on subsequent competitions are positive. Table A8 shows the baseline results without individual fixed effects and they are indeed larger.

[28]While one might consider this negative signaling effect around the cutoff as a source for bias that increases the effect sizes, I argue that the negative slopes are not particularly large.

Figure 1.4: RDD plots - Effects on team formation



Note: Binned scatterplots with local linear regression. The p-value of the discontinuity estimate is displayed in the top left corner. The size of the circles shows the relative average weight of observations within each bin. Weights are given by the triangular kernel function and decrease linearly with the distance from the cutoff. Only the area of 2.5% of the leaderboard on both sides of the cutoff are taken into account for the regression and displayed here. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. Covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code.

believe in the effectiveness outside of the platform as well. To test this hypothesis, I next look at whether users use the signal on their digital résumés.

1.6.2 Effect on signaling activity

Here, the results from the second research question are reported to find out whether winning a medal leads to different signaling patterns on LinkedIn. Table 1.3 panel a) reports the OLS results: Column 1 shows that competitors who obtained a medal have a 2.7 percentage point higher likelihood of providing a link to their LinkedIn page on their competition profile and this represents a 9% increase from the sample mean. Additionally, winners are 9 percentage points more likely to mention the name of the competition website on their LinkedIn profiles, as shown in column 2. Columns 3 to 5 indicate whether comparable but competition-unrelated forms of signaling are positively or negatively related to medals. These observed signals are counts of the number of certificates, recommendations, and skills listed, all of which have negative coefficients. Every effect measured by OLS is highly significant, with the exception of the number of skills which is

Table 1.3: Estimation results: Effects on signaling

	Comp. signals		Oth. signals		
	(1) Links LinkedIn	(2) Mention competitions	(3) Certificates	(4) Recommendations	(5) Skills
a) OLS:					
Medal winner	0.02680***	0.09039***	-0.87325***	-0.37109***	-0.47879**
	(0.00311)	(0.00715)	(0.11918)	(0.05088)	(0.20588)
Observations	158331	22493	22493	22493	22493
Individuals	53371	2859	2859	2859	2859
b) RDD:					
Medal winner	0.04410**	0.04392	-0.57776	-0.75949*	-0.65079
	(0.01900)	(0.05502)	(0.62166)	(0.45605)	(1.40089)
Observations	9709	1686	1686	1686	1686
Individuals	8790	1450	1450	1450	1450
c) IV:					
Medal winner	0.07036***	0.22325**	0.09979	0.17832	-1.15382
	(0.02321)	(0.10597)	(1.24861)	(0.63062)	(2.70891)
Observations	9445	1103	1103	1103	1103
Individuals	8637	935	935	935	935
1st stage F	1607.63	102.99	102.99	102.99	102.99

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

significant on the 5% level. To find out whether this really indicates that competitors use the signal and substitute away from other forms of signaling, we need to turn to methods that allow for causal interpretation.

The RDD estimates in table 1.3 panel b) show that the effect for providing a link to LinkedIn in column 1 is significant on the 5 percent level and represents an increase by almost 30%. Other than that, only the effect on the number of recommendations is significant on the 10 percent level. The IV estimates are reported in the bottom panel of table 1.3.[29] Winning a medal causes a highly significant 7 percentage point increase in the likelihood of linking to one's professional profile (column 1) and a significant and large effect of 22 percentage points on the likelihood of mentioning the competition website (column 2). The non-competition related signals have no significant effects.

These results indicate that there is some evidence that winning a medal influences signaling behavior but there is no clear evidence that participants substitute away from other signals. I interpret these findings that competitors do believe they could benefit professionally from their competition achievements.

1.6.3 Effectiveness of the signal on the labor market

Ultimately, to see whether innovation tournaments help aspiring data scientists to demonstrate their skill, we need to know if medal winners are in a better position on the labor market. An indicator for holding a data science position shall tell us whether the signal facilitates entry into this interdisciplinary field. As described in section 1.5.2.2, estimates can be biased toward zero if competitors use medals as a substitute for other signals. Even though table 1.3 showed that there is no causal substitution away from these, certificates, recommendations and the number of skills are still included as controls in the estimations.[30] Table 1.4 reports the results for the baseline specification, the RDD, and the IV. The baseline result in panel a) is significant on the 1% level, with the direction as hypothesized.

The RDD estimates reported in table 1.4 panel b) show no significant effect but the coefficient is comparable in size. The regression discontinuity plot can be found in figure A7 and it appears to show null effects as there is no visible discontinuity. Table 1.4 panel c) reports the results from the IV estimation. The point estimate also points in the same direction as the OLS coefficient but the effect is not statistically significant.[31]

1.6.4 Field of study heterogeneity

We can get a clearer image into the results by looking at treatment effect differences by field of study. The original hypothesis was that an easy to screen signal can help aspiring data scientists to enter a young and interdisciplinary field especially when they transition from various backgrounds. However, when they can already prove their expertise, the value of a medal as a signal may be reduced, thus eliminating the need for it. A degree in one of the recent Data Science programs or in the two most related traditional fields,

[29] More detailed results for the IV, including the first stage, can be found in appendix table A6.

[30] Results without these specifications can be found in the appendix table A9. The results are roughly comparable, with point estimates even slightly smaller in absolute size for the OLS and RDD.

[31] More details on the first stage can be found in appendix table A7.

Table 1.4: Estimation results: Effects on labor market success

	(1) Data Scientist
a) OLS:	
Medal winner	0.02252***
	(0.00852)
Observations	22493
Individuals	2859
b) RDD:	
Medal winner	0.01909
	(0.05651)
Observations	1686
Individuals	1450
c) IV:	
Medal winner	0.03959
	(0.10832)
Observations	1103
Individuals	935
1st stage F	103.40

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. All specifications include competition fixed effects.

Computer Science and Statistics, can be considered such a signal. Hence, I analyze both signaling behavior and labor market success for individuals with and without such a degree. Information about degrees is taken from participants' LinkedIn profiles[32]

Results are reported in table 1.5. The OLS results in panel a) show that, after winning a medal, holders of a related degree are significantly less frequently mentioning the competition website (column 1) and less frequently working as data scientists (column 2) than competitors without such a degree. Where the RDD effect on the likelihood of mentioning competitions on LinkedIn was previously not statistically significant, we can see in column (1) of panel b) that, for each group individually, the effect is highly significant but goes in opposite directions. A similar pattern, even if not significant, can be seen for the likelihood of working as a Data Scientist in column (2). The regression discontinuity plots in figure 1.5 also visualize this finding clearly: Medals decrease the likelihood of using the signal for participants from related fields but increase it for participants from other fields. This suggests that individuals consider medals to be much more valuable when they do not already have a more meaningful way of showing their knowledge. At the same time, it appears that for individuals that already hold a degree in a related field, the medal is even considered to be a negative signal. It is possible that the effect on the likelihood of working as a data scientist is statistically insignificant due to a lack of statistical power since the point estimates are relatively large. The results from the IV estimation in panel c) are insignificant and do not fully confirm the findings from the other methods: Here, the coefficient of the interaction is positive as well.

A possible reason for why the RDD shows that medals appear to be a negative signal for holders of a related degree and why this is not observed with the IV method, is that the RDD exclusively considers bronze medals, whereas the IV effect is to some extent also identified from silver and gold medals as has been discussed in section 1.5.1.

1.6.5 Sensitivity analyses

In this section, I want to show whether the results change under different conditions such as alternative specifications, and address possible concerns.

Silver and Gold As mentioned before, it is possible that the effects are stronger for silver and gold medals than for bronze. To test this, I estimate the RDD and IV effects for silver and gold medals and report the results for the three research questions respectively in tables A10, A11, and A12.[33] For the team formation outcomes, silver and gold medals do not lead to better team formation possibilities when estimating with the RDD, as shown in table A10. However, the IV estimates in panels c) and d) do show stronger effect sizes for higher achievements. One possibility for this finding is that while the RDD strictly estimates the effect of an incrementally better medal, the IV additionally captures the effect of a silver or gold medal relative to none at all. This is the result from compliers with score changes large enough to net them a gold medal where they would

[32]For this reason, the outcome *linked LinkedIn* cannot be included in this analysis.

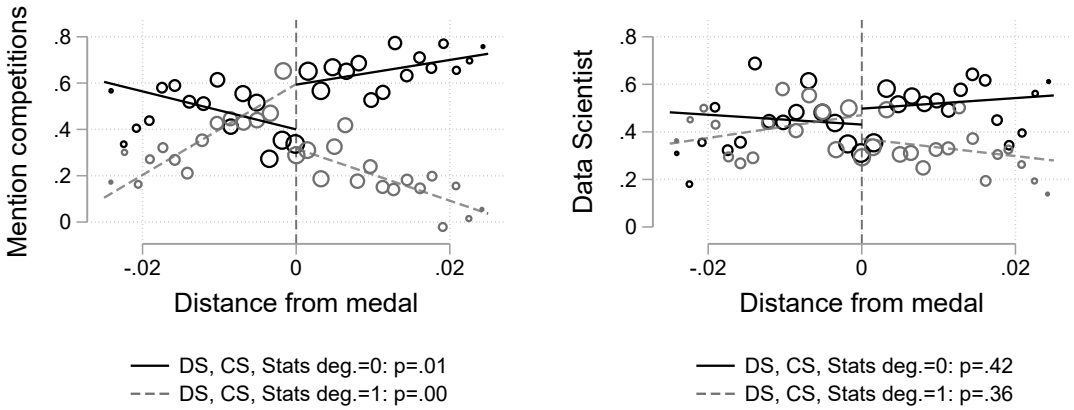
[33]This has not been planned at the time of data collection, where only profiles from individuals around the bronze cutoff and a random sample of IV compliers for any medal have been sampled. Thus, for the signaling activity and labor market success, the effects of silver and gold medals can only reliably be estimated with the IV method.

Table 1.5: Estimation results: Heterogeneity with respect to degree

	(1) Mention competitions	(2) Data Scientist
a) OLS:		
Medal winner	0.10234***	0.04768***
	(0.00936)	(0.01132)
DS, CS, Stats deg.	0.00768	-0.01037
	(0.00747)	(0.00805)
Medal winner × DS, CS, Stats deg.	-0.02229*	-0.05659***
	(0.01174)	(0.01412)
Observations	22493	22493
Individuals	2859	2859
b) RDD:		
Medal winner (DS, CS, Stats deg. = 0)	0.19338**	0.06692
	(0.07601)	(0.08349)
Medal winner (DS, CS, Stats deg. = 1)	-0.27854***	-0.09962
	(0.10411)	(0.11016)
Observations	1686	1686
Individuals	1450	1450
c) IV:		
Medal winner	0.24705	0.22882
	(0.22135)	(0.22726)
Medal winner × DS, CS, Stats deg.	-0.04544	-0.31336
	(0.32965)	(0.33846)
DS, CS, Stats deg.	0.06017	0.20151
	(0.18719)	(0.19219)
Observations	1103	1103
Individuals	935	935
1st stage F	18.10	18.10

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

Figure 1.5: RDD Plots - Heterogenous effects by field of study



Note: Binned scatterplots with local linear regression. The p-values of the discontinuity estimates are displayed in the legend. The size of the circles shows the relative average weight of observations within each bin. Weights are given by the triangular kernel function and decrease linearly with the distance from the cutoff. Only the area of 2.5% of the leaderboard on both sides of the cutoff are taken into account for the regression and displayed here. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. Covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. Includes competition fixed effects.

otherwise not have gotten any. Likewise, table A11 shows that higher medals lead to stronger effect sizes for *linking LinkedIn*, *mentioning the competitions* and labor market success in table A12 but the latter remains insignificant.

Continuation with different team For the outcome *team participation* in the team formation analysis, one might argue that after winning a medal, a team is more likely to stick together, increasing the likelihood of subsequent team participation. To correct for this, I estimate a specification that explicitly rules this case out, i.e., ignores observations where the team in the outcome competition is the same as in the treatment competition. The results in table A13 column 1 show that the effect sizes are indeed slightly smaller and no longer significant for the IV method but they remain significant for the RDD.

Exclusion of pure solo participants It is possible that some competitors have no interest in team participation and compete by themselves exclusively. For these, we should not expect any effect of the signal on team formation. Table A14 reports the results when only participants that ever participated in any team with others before are included. As expected, the coefficients for both the OLS and the RDD specification are stronger than for the full sample.

Correction for multiple treatments A drawback of the research questions involving time invariant outcomes is that some competitors appear in the treatment or control groups more than once and yet they are handled as independent observations. This could introduce bias which I address by including only individuals that appear in the natural experiment as compliers once. For the RDD, this means competitors that appear within the 2.5% leaderboard window in more than one competition are dropped. For the IV, only those are considered whose medal status changed due to a score change a maximum of one time. The results reported in tables A15 and A16 show that the estimates do not substantially change. For *linked LinkedIn* in the first column of table A15, the results are significant for the RDD (panel a) and stay significant in the IV specification (panel b). Referring to the competitions as outcome in column 2 is no longer significant for the reduced sample specification in panel b), however, this is due to the larger standard errors as the point estimate is slightly larger then before. The effect on labor market success in table A16 remains insignificant.

Controls for indirect effect My approach does not isolate the direct signaling effect. In particular, it is possible that the achievement in one competition leads to better team formation and thus better overall performance in subsequent competitions and this affects signaling decisions and career opportunities. Hence, the estimated overall effect would be larger than the direct signaling effect. To account for this, I add additional controls for the number of medals won *after* the treatment competition. Tables A17 and A18 show the results of this exercise. As expected, the coefficients are slightly smaller but go in the same direction as in the main specification and for the most part, the effects remain statistically significant as long as they are in the main specifications.[34]

[34]The only exception is the coefficient on the number of recommendations for the RDD.

Heckman sample selection correction Furthermore, one might argue that there could be sample selection bias for both the within-competition analysis as well as the research questions involving LinkedIn data as outcomes. For the team formation analysis, identification can only stem from individuals that continue participating after the treatment competition. However, winning a medal can influence the decision to continue participation. And indeed, table A19 suggests that winning a medal increases the likelihood of continued participation. For the outcomes from online résumés, there can be selection in terms of who is and who is not on LinkedIn and also provides a link to the platform. Hence, I use the procedure by Heckman (1979) to correct for sample selection bias. For the first research question, i.e., the effects on team formation, I model *continuing participation* as a function of winning a medal, information about the person from their profile[35], and other covariates. For the research questions involving outcomes from LinkedIn, I model *Linked LinkedIn* as a function of winning a medal, other covariates, whether also the website *Github* and a private URL have been linked, and the number of followers. Unfortunately, I can only correct for selection with respect to *linking* an account, not for *having* one because I cannot observe this. Tables A20, A21, and A22 respectively show the results for the three research questions. Overall, the results are no longer significant for the team formation outcomes and the RDD estimate for mentioning competitions.

Different functional forms A general concern is that Regression Discontinuity Designs can be sensitive to specifications of the functional form, such as the use of the specific kernel function that is weighting observations by their distance to the threshold. Tables A23, A24, and A25 show the results for different kernel functions for all research questions respectively. While the effect sizes do vary, the results remain unaffected in direction and the (in)significant ones remain (in)significant.

Different bandwidthes The RDD estimates could also be too large if the relation between the running variable and the outcome is in fact not discontinuous but rather non-linear such that the jump of the fitted lines is primarily driven by points further away from the threshold. To account for this, figures A8, A9, and A10 show how the estimates change when differently sized windows around the threshold are considered. If the effects were driven by nonlinearity, the effects should disappear for smaller bandwidthes. The graphs show that the conclusions do not substantially change. If anything, most estimates appear to be even more pronounced with a reduced bandwidth. Only few switch signs while the effects remain insignificant and if so, this happens only for those outcomes that have insignificant results to begin with.

1.7 Discussion

A potential explanation for the lack of strong findings lies in the possibility of power concerns in the analyses with resumé based outcomes. For example, figure 1.5 showed a sizeable discontinuity in the probability of working as a Data Scientist for people from not

[35] i.e. dummies for whether they provided links to their profiles on other websites and the number of their followers.

immediately related backgrounds but the results are nonetheless insignificant. However, the point estimates are still substantially smaller than those for the significant signaling outcome.

The effects on labor market success may also be insignificant because I estimate Local Average Treatment Effects. In particular, the effect of winning a bronze medal. Even if it still indicates that one is an above average competitor, sometimes in the top 10%, a bronze medal might have a negative connotation. However, at least among participants, bronze medals are valued positively since they do have a positive effect on team formation among peers and make participants more likely to connect their competition and professional profiles. At the same time, the effects of a silver or gold medal on peer recognition are mixed as the Regression Discontinuity analysis shows no additional advantage relative to a bronze medal. Hence, for bronze to be considered negatively, employers and recruiters would have to think about this achievement very differently than competitors.

A shortcoming of the analysis on effects on labor market success is that it is not perfectly observable whether and to what extent the medal has been used as a signal. Also, it is not clear for what purpose participants use LinkedIn, i.e., whether they use it to increase their visibility or whether they use it only to inform themselves. Thus, the estimated effects only reflect Intention to Treat Effects.

In terms of external validity, it is not clear how well this analysis translates to other fields: Because I estimate effects of placing relatively high but not at the top, the results may only hold for competitions where it is possible to make out a relative position, i.e., ideally with an objective ranking system. Something like this is not straightforward to implement for example for graphic design contests with one or few winners and no ranking of losers. Additionally, the results may not hold over time since it is possible that new technologies further influence how employers select candidates and where they receive their information from.

Here, different related professions that apply predictive modelling, such as *Machine Learning Engineer*, are all classified manually and subsumed under the "Data Scientist" label. Not only does this introduce the possibility for some measurement error, but because these professions have different tasks, they may benefit differently from such a signal. Furthermore, competitions measure only part of the skills demanded by employers: Running and optimizing machine learning models makes up only a fraction of the work of a typical data scientist. Formulating the problem statement, data acquisition and cleaning is often already done by the platform and there is no need to communicate the results.

1.8 Conclusion

The internet offers new opportunities for workers to show what they are worth and thousands in the data and software professions are displaying their skill via different online channels. However, it is not always clear whether they are motivated by the opening up of potential new job opportunities and whether this is even effective.

This paper studies the labor market signaling effect of skill indicators from online innovation tournaments for Data Science where public competition placement can be an informative indicator for skill. Using rich data from a leading platform in this field, I

estimate whether digital achievement tokens serve as an effective signal both within and outside of the platform. I find evidence that competition success does translate into a higher likelihood of joining a team in the following competition. Furthermore, winning a medal increases the likelihood that participants link their professional and competition profiles by both providing a link to their CVs on their competition profiles and by mentioning the competitions on their CVs. An interpretation for this is that competitors do think the achievements function as a labor market signal to some degree. However, the effect on the actual likelihood of entering the field of Data Science professionally goes in the expected direction but is not significant. A heterogeneity analysis suggests that participants who do not have an academic background in a field most related to data science consider the medal to be more valuable.

Future research may address the limitation of bronze medal effects and instead estimate the effect of stronger signals from online competitions. Extensions are possible by further analysing participants' motivations to compete or by looking more into heterogeneous effects to see whether potentially disadvantaged groups, such as migrants or women, which appear to be in the minority, benefit more from the signal. It is also interesting to go beyond the effects on participants on this platform and either compare it to other platforms and different fields or evaluate whether such competitions are worthwhile for the companies posing the challenges.

The labor market matching process of the future may be very different from what we are used to. Already, algorithms aid employers in the selection process, at the very least in online labor markets (Horton, 2017). State of the art recommender systems would be capable to incorporate a plethora of features from various publicly or otherwise available data sources. It is easy to imagine that verifiable platform credentials like achievements on tournament websites can serve as a useful variable in decision making. As of yet, however, these tournaments help aspiring data scientist most by offering an open community to learn and share ideas.

Chapter 2

Gender Stereotypes in User Generated Content*

2.1 Introduction

Despite advances during the past decades, important hurdles remain on the path to gender equality. In particular, gender stereotypes persist (Bertrand, 2020). Gender stereotypes reflect general expectations about attributes, characteristics, and roles of women and men. E.g., assertiveness and performance are often ascribed to men, while warmth and care for others are attributed to women (e.g., Kite et al., 2008; Fiske, 2010). Recent empirical evidence demonstrates that gender stereotypes affect how we perceive others and how we perceive ourselves (Ellemers, 2018), confining both personal choices and professional careers (Jensen and Oster, 2009; La Ferrara et al., 2012; Kearney and Levine, 2015). Thus, assessing and addressing gender stereotypes in our society is of utmost importance.

How prevalent are gender stereotypes? It is difficult to address this question, as stereotypical beliefs are not always conscious, and even if they are, they may not be openly expressed (Blackburn, 2017).[1] The growing importance of user-generated content (UGC) opens up novel opportunities to overcome such biases, though. In particular, the anonymity of users in online discussion fora may eliminate social pressures and allow individuals to voice what they think but would otherwise not say (Hsueh et al., 2015; Wu, 2018). At the same time, recent developments in automated text analysis (Gentzkow et al., 2019; Ash and Hansen, 2022) provide the necessary tools to assess gender stereotypes in UGC at large-scale.

This paper leverages a unique dataset of more than a million anonymous comments from a major German online discussion forum to examine the prevalence and development of gender stereotypes over time. To this end, we combine several state-of-the-art text analysis and machine learning techniques that classify (i) whether a comment discusses men or women (or no person at all), and (ii) whether a comment covers topics that are stereotypical male (related to work and money) or stereotypical female (related to family, home, and physical appearance) (Fiske, 2010; Ellemers, 2018; Marjanovic et al., 2022). Taken together, the gender and topic classifications allow us to assess if men are

*This chapter is based on joint work with Anna Kerkhof.

[1] E.g., social desirability bias – the tendency to provide answers that adhere to social norms – is likely to confound self-reported measures (Podsakoff, 2003; Fisher, 1993).

mentioned more often than women in the context of "male", and women more often than men in the context of "female" topics at a given point in time. Based on that, we can document whether, where, and to what extent gender stereotypes exist in our data, and how they develop over time.

The topic classification of comments is conceptually challenging, though. In particular, we wish to assess which topics are being discussed such that the inference is not driven by gender itself. E.g., a classic supervised machine learning (ML) algorithm could learn patterns like "Comment talks about women, thus higher likelihood of topic *family*" from the training data and transfer them to the sample of interest. As a result, we would not be able to detect differences in gender stereotypes between the training and the prediction sample and, crucially, we would not be able to detect changes in gender stereotypes over time. Dictionary methods that use curated lists of words related to specific topics could address this issue. However, classic dictionary methods are prone to yield both false positives and false negatives, and they are sensitive to prefixes, suffixes, synonyms, and typographical errors.

We propose an innovative solution to these challenges by enriching unbiased dictionaries with the flexibility and "understanding" of *word embeddings* (Mikolov et al., 2013). Word embeddings represent the semantic meaning of words in an n-dimensional space, where the embedding vectors of words with similar meaning are close to each other. We exploit this feature by transforming words associated to specific (gender stereotypical) topics – e.g., work or family – from a dictionary into their word embedding representation.[2] Then, we generate a large number of linear combinations of the word embeddings associated to one specific topic, where the resulting vectors lie somewhere in between the original embeddings. Under the key assumption that word embeddings associated to specific topics are clustered in the vector space, we can use these linear combinations as unbiased training data for a supervised ML algorithm (Support Vector Machine) that is ultimately able to predict if a particular comment covers a specific topic or not.

To apply the trained model to our sample of interest, we must make multi-word comments comparable to word-level embeddings. To this end, we determine each comment's most important words through a *clustered tf-idf* approach. Then, we transform these words into their word embedding representation and compute their linear combination, using their normalized *tf-idf*-values as weights. Each comment is thus ultimately represented by a linear combination of word embeddings that is projected onto the same vector space as our training data, whereby we can apply the trained model for topic classification.

In contrast to the more ambiguous (gender stereotypical) topics, the occurrence of men and women as part of the discussion in our comments is relatively explicit. As a result, we can base our gender classification on a composite of classic dictionary approaches. To minimize the number of false positives, we restrict the procedure to carefully selected gender specific names and terms. To minimize the number of false negatives, we combine three different dictionary approaches that complement each other.

Based on our topic and gender classification, we present strong evidence for the prevalence and persistence of gender stereotypes in our data. In particular, we show that men are relatively more often discussed in the context of "male" topics like work and money

[2]Specifically, we use the *Linguistic Inquiry and Word Count Dictionaries* ("LIWC"); see Section 2.3.2 for details.

than women, and women are relatively more often discussed in the context of "female" topics like family, home, and physical appearance than men. Moreover, while gender stereotypes related to work, money, and physical appearance slightly diminish over time, we find no such pattern for domestic issues like family and home. These findings are further supported by regression analyses that control for comment characteristics as well as user and news section fixed effects. The results are robust to excluding offensive language from our data, and they are not driven by potential stereotypes in the news articles that the comments were originally attached to.

Researchers have recently started to distinguish between hostile and benevolent sexism (e.g., Glick and Fiske, 2001, 2018). While both are based on gender stereotypes, hostile sexism conveys a clear antipathy, whereas benevolent sexism is positive in tone but imparts patronizing beliefs about women.[3] To examine whether the gender stereotypes in our data are driven by hostile or benevolent sexism, we first determine their sentiment, and then use standardized sentiment scores as weights for our comments. In line with our analysis of offensive language, we find just small evidence for the existence of benevolent sexism in the context of work, money, and physical appearance, and no evidence for either benevolent or hostile sexism in the context of domestic issues.

Our paper makes two major contributions to the literature. First, we advance the broad and timely research on gender inequality and gender discrimination (e.g., Bertrand and Duflo, 2017). As far as we know, we are the first who leverage the anonymity of UGC to provide a clean and extensive analysis of the prevalence and development of gender stereotypes over almost a decade. Second, we develop a novel ML-based procedure to classify UGC, where we enrich classic dictionaries with the flexibility and understanding of word embeddings. This procedure can be used more generally for document-level topic classification; potential applications include all types of novel and unconventional text as data such as social media and other online platforms.[4] To further support research in that direction, our method is available as a Python package on `https://github.com/VFMR/WEELex`.[5]

The remainder of the paper is organized as follows. Section 2.2 reviews the related literature on gender discrimination and stereotypes, UGC, as well as on recent advances in automated text analysis. Section 2.3 describes our data and illustrates both the topic and the gender classification in detail. In Section 2.4, we apply these classifications to our data and illustrate the prevalence and development of gender stereotypes over time. Section 2.5 provides a battery of robustness checks on our main results. Section 2.6 concludes.

2.2 Related literature

Our paper is related to three strands of literature. First, we add to the vast research on gender inequality, in particular to studies on gender discrimination (e.g., Altonji and

[3]E.g., a man's comment to a female colleague on how "cute" she looks, however well-intentioned, may undermine her feelings of being taken seriously as a professional (see Glick and Fiske, 2018).

[4]See Section 2.6 for further discussion.

[5]More specifically, our package supports document-level classification with independent categories as well as polarity detection using a dictionary of weighted terms via an implementation of Latent Semantic Scaling (Watanabe, 2021).

Blank, 1999; Blau and Kahn, 2017; Charles and Guryan, 2011; Bohnet, 2016; Bertrand and Duflo, 2017) and gender norms and stereotypes (e.g., Akerlof and Kranton, 2000; Bordalo et al., 2016, 2019; Ellemers, 2018; Bertrand, 2020; Ash et al., 2021a,b). Most of this literature considers gender discrimination in specific contexts (e.g., in the work place) or discusses the prevalence of gender stereotypes at a given point in time. We contribute by examining the prevalence and development of gender stereotypes in UGC over the course of almost a decade, where the anonymity of users allows us to overcome unconscious and social desirability biases that often confound self-reported measures. Moreover, despite the growing importance of online discussions, gender stereotypes in UGC have hardly been studied before.

Most closely related are the papers by Wu (2018) and Marjanovic et al. (2022). Wu (2018) studies the prevalence of gender stereotypes in the "Econ Job Market Rumors" forum and finds that the discourse becomes significantly less academic oriented, and more about personal information and physical appearance, when users talk about female researchers. Relatedly, Marjanovic et al. (2022) examine gender stereotypes in about ten million comments on male and female politicians from Reddit and show that female politicians are more often described in relation to their body, clothing, and family than males. We extend these analyses in three important ways. First, we analyze an extensive amount of comments on a broad range of topics from a general interest discussion forum, which enhances the external validity of our results compared to the existing studies. In particular, our findings are not limited to gender stereotypes held by a subset of economists, or gender stereotypes related to politicians.[6] Second, both Wu (2018) and Marjanovic et al. (2022) provide static analyses, while our paper examines the prevalence and development of gender stereotypes over time. Third, in contrast to our study, neither of them addresses potential gender biases in the topic classification of UGC.

The second strand of related literature examines UGC (see Luca, 2016, for a survey). The lion's share of this research focuses on the analysis of consumer reviews (e.g., Chevalier and Mayzlin, 2006; Mayzlin et al., 2014; Anderson and Magruder, 2012) or incentives to contribute UGC (e.g., Wang, 2010; Anderson et al., 2013; Easley and Ghosh, 2013; Zhang and Zhu, 2011). While text analysis – especially sentiment analysis – is not new to this literature, UGC has thus far not been tapped to examine the prevalence and, in particular, the development of large societal phenomena such as gender stereotypes. Moreover, the anonymity of users has rarely been considered as a feature, but rather as a problem, e.g. in the context of hate speech (Gagliardone et al., 2015).

Third, we propose a new procedure to classify UGC and thereby add to the growing research on text as data (Grimmer and Stewart, 2013; Gentzkow et al., 2019; Ash and Hansen, 2022). The novelty of our approach is to enrich classic dictionary methods with the flexibility and understanding of word embeddings as developed by Mikolov et al. (2013) and Bojanowski et al. (2017). We thereby contribute to a vibrant literature that incorporates NLP and ML methods to answer economic questions that could not be addressed before (Athey, 2019; Athey and Imbens, 2019). Our paper is especially close to Garg et al. (2018), who use word embeddings to quantify historical trends and social

[6]The readership of *Spiegel Online* is predominantly male, middle-aged, well educated, and well earning; see `https://app.powerbi.com/` for the most recent readership data collected by the Working Group on Media Analysis (*Arbeitsgemeinschaft Media-Analyse e.V.*, homepage: `https://www.agma-mmc.de/`).

change in gender and ethnic stereotypes. However, while Garg et al. (2018) explicitly allow their word embeddings to capture gender stereotypes, our approach is especially designed to prevent this. In addition, most of the literature on text as data studies English corpora, while analyses involving other languages are rare. We contribute to closing this gap by developing a classification procedure that we apply to German data, but which could principally be used for all languages that feature appropriate (unbiased) dictionaries and pre-trained word embeddings.[7]

Our classification procedure as such is furthermore related to two recent sub-strands of research in text analysis. First, it links to *Correlation Explanation (CorEx) Topic Modelling* (Gallagher et al., 2017), an anchored topic modelling approach using seed words – i.e., a dictionary – to assign documents to topics. This method uses the entire corpus to determine the best fitting topics, though, whereby it is susceptible to issues of gender bias as described above. Second, our approach is similar to *Latent Semantic Scaling* (Watanabe, 2021), which combines a dictionary with word embeddings, too, but is limited to predictions along a single axis (e.g., a sentiment score or political polarity). Likewise, the *Word Embedding Association Test* by Caliskan et al. (2017) employs word embeddings to measure the similarity of words to predefined topics, but also operates on just one dimension. We add to this literature by developing a topic classification procedure that avoids gender bias and is furthermore able to predict multiple topics that are not mutually exclusive.

2.3 Data and classification

Our analysis of gender stereotypes in UGC features a unique sample of about 7.5 million comments that we classify through an innovative combination of classic dictionary methods, word embeddings, and supervised ML algorithms. This section illustrates the raw data and describes our topic, gender, and sentiment classification procedures in detail.

2.3.1 Data

Our data comprises 7,345,166 comments that we retrieved from the public *Spiegel Online* (*"SPON"*) discussion forum by the end of 2019.[8] *SPON* attracts around 19 million users per month[9] and ranks among Germany's top five online news websites.[10]

SPON allows its users to comment and discuss its news content. The comments are organized in threads that are attached to *Spiegel Online*'s news articles, but the discussion could also be accessed through a central interface that aggregates all threads. Around 70% of all news articles allow for comments; the remaining 30% typically involve sensitive issues such as migration, terror attacks, and sexual harassment (Dachwitz, 2016).

[7]Alternatively, if pre-trained word embeddings do not exist, a sufficient requirement is to leverage a corpus large enough to train one's own embedding vectors.

[8]Since Jan 2020, users must log in to the forum to read and write comments, which eliminates the anonymity that we wish to exploit for our analysis.

[9]See `https://meedia.de/2017/04/13/agof-welt-rueckt-dank-n24-traffic-an-spon-heran-focus-dank-rekordzahlen-fast-gleichauf-mit-bild/` (Dec 2022).

[10]See IVW, `https://ausweisung.ivw-online.de/index.php?it=1&setc=1` (Dec 2022).

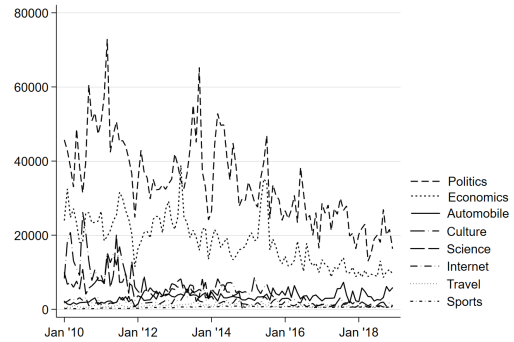
For each comment, we retrieve information on the user alias (i.e., the nickname of the user who has written the comment), the time and date of upload, position in the thread, and the content of the comment itself. Note that we cannot infer the users' gender from their aliases, and that individual comments usually do not explicitly refer or respond to previous comments from the same thread. Appendix B.1.1 displays some exemplary comments, Appendix B.2 illustrates one exemplary discussion thread in detail.

Figures 2.1a to 2.1f describe our raw data in more detail. Figure 2.1a depicts the absolute number of comments posted within each of *SPON*'s news sections from Jan 2010 to Dec 2018. Plausibly, the majority of comments is attached to articles on politics or economics, which are *SPON*'s most important news sections. While the absolute number of comments per month is impressive (e.g., 126,990 comments were posted just in March 2011), Figure 2.1a also reveals that it has been shrinking over time. However, Figure 2.1b shows that part of the effect can be explained by a diminishing number of articles that allow for discussion on behalf of the users, especially after the 2015 refugee crisis (we observe a total of 782,431 articles/threads). There is ample heterogeneity in the number of comments per thread: while the median (mean) thread features 10 (9.43) comments, the minimum number is equal to 1 and the maximum number equal to 80. Similarly, the comments' average length varies a lot, with a median (mean) length of 336 (467.81), and a maximum length of 23,239 characters (Figure 2.1c).

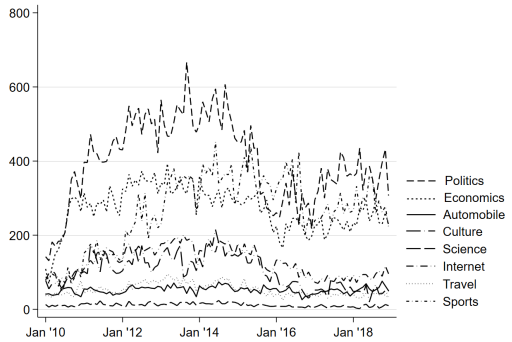
Considering the users ($n = 272,023$), we find that the majority of comments is written by a minority of users. E.g., the median user posts just two comments, the mean user 27, and the most active users several thousands (Figure 2.1d). While some users just post one comment and never come back again, others remain active for considerable time periods. In particular, we find that the minimum amount of time between the first and the last comment is equal to zero for the majority of users, but that there is a long tail of users who remain active for several years (Figure 2.1e). The users are not too specialized in terms of topics that they contribute to. Specifically, Figure 2.1f shows that, conditional on writing at least two comments, many of them contribute to discussions related to two or more news sections.

To examine the content of the comments in further detail, we use BERTopic (Grootendorst, 2022), a state-of-the-art NLP topic modeling technique to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. To focus on the most important aspects, we restrict the analysis to comments that we eventually classify as discussing men or women (i.e., that we classify as *male* or *female*, respectively).

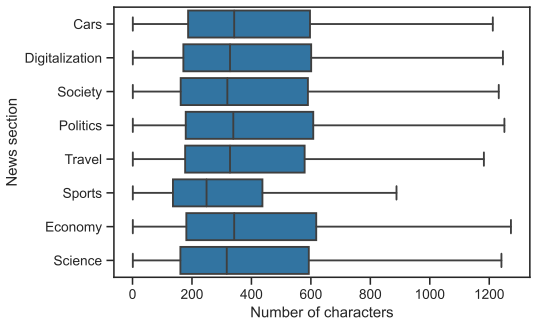
Figure 2.2a displays the most important terms for the most important topics in comments classified as *male* or *female*, respectively. We find that political issues prevail; in particular, an immense proportion of comments classified as *female* seems to be about Angela Merkel. Exluding such comments from the analysis (Figure 2.2b) reveals that many comments about women discuss gender related issues such as sexism, feminism, and leadership quotas. We also find that the relevance of the topics varies over time; e.g., Figure 2.3 shows that the financial crisis in Greece was a major topic in 2015 and that debates on muslims and kurds domineered in 2016, shortly after the infamous terror attacks in Paris. Similarly, we find that debates on sexism gained importance with the #MeToo movement in 2018.



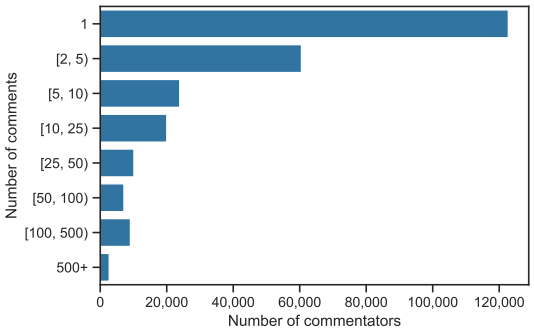
(a) Absolute number of comments per news section



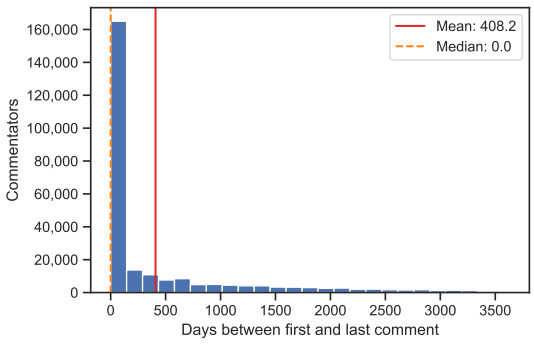
(b) Absolute number of threads per news section



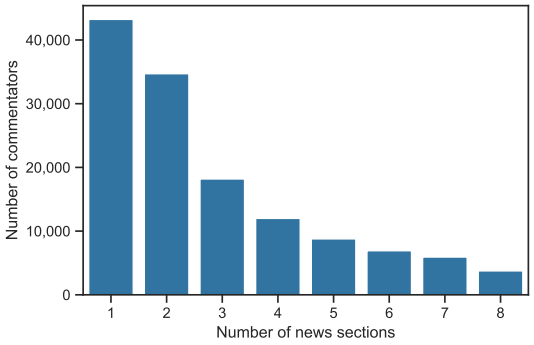
(c) Boxplot number of characters per comment by news section (no outliers).



(d) Distribution of the number of comments per user.

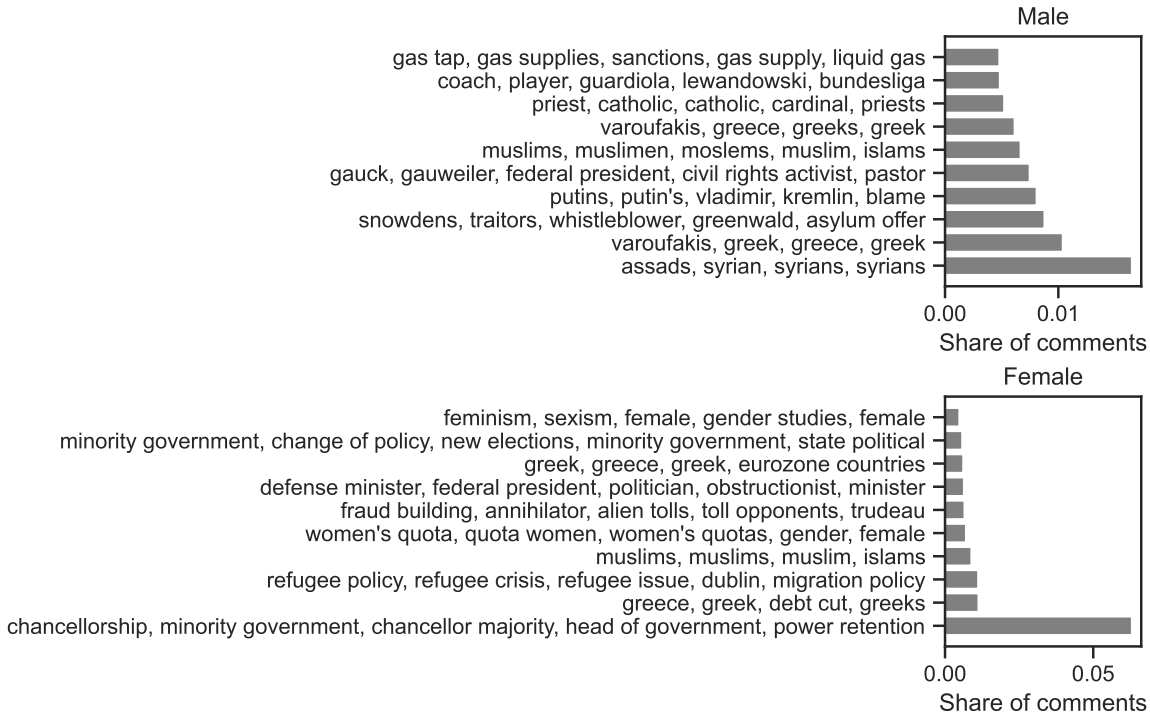


(e) Activity of users.

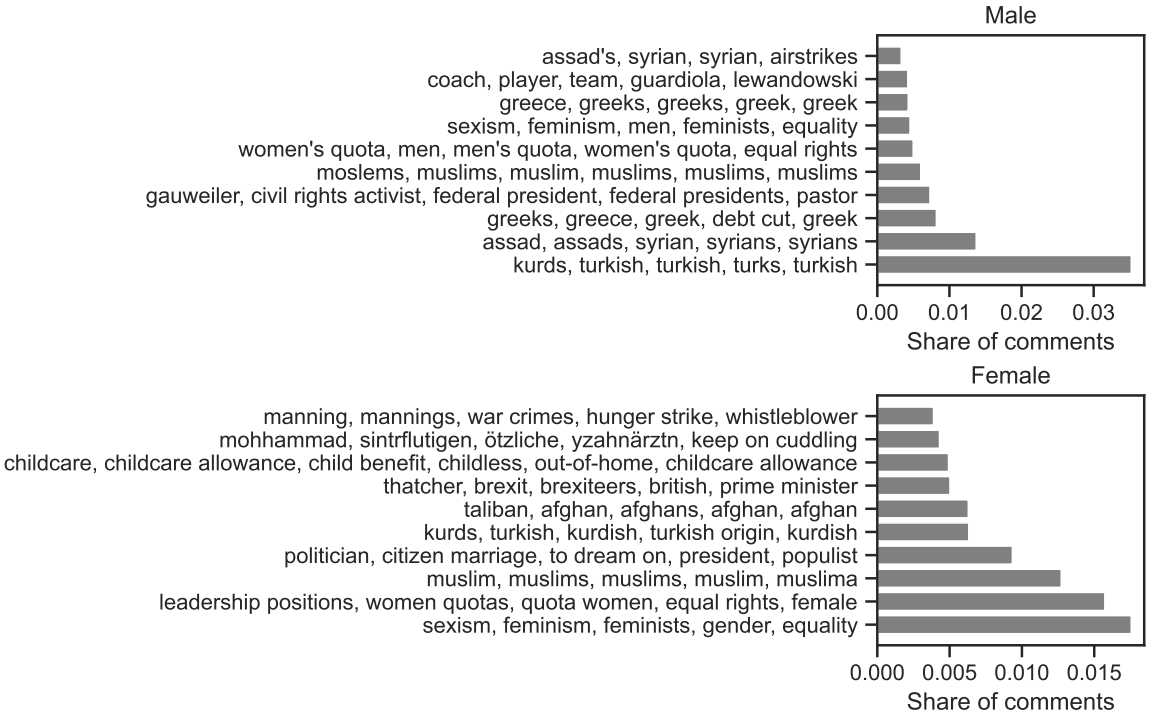


(f) Number of different news sections that a user contributes to, for users with at least two contributions.

Figure 2.1: Descriptives of the raw data.



(a) Most important terms of the most frequently occurring topics.



(b) Most important terms of the most frequently occurring topics, without comments on Angela Merkel.

Figure 2.2: BERTopic output.

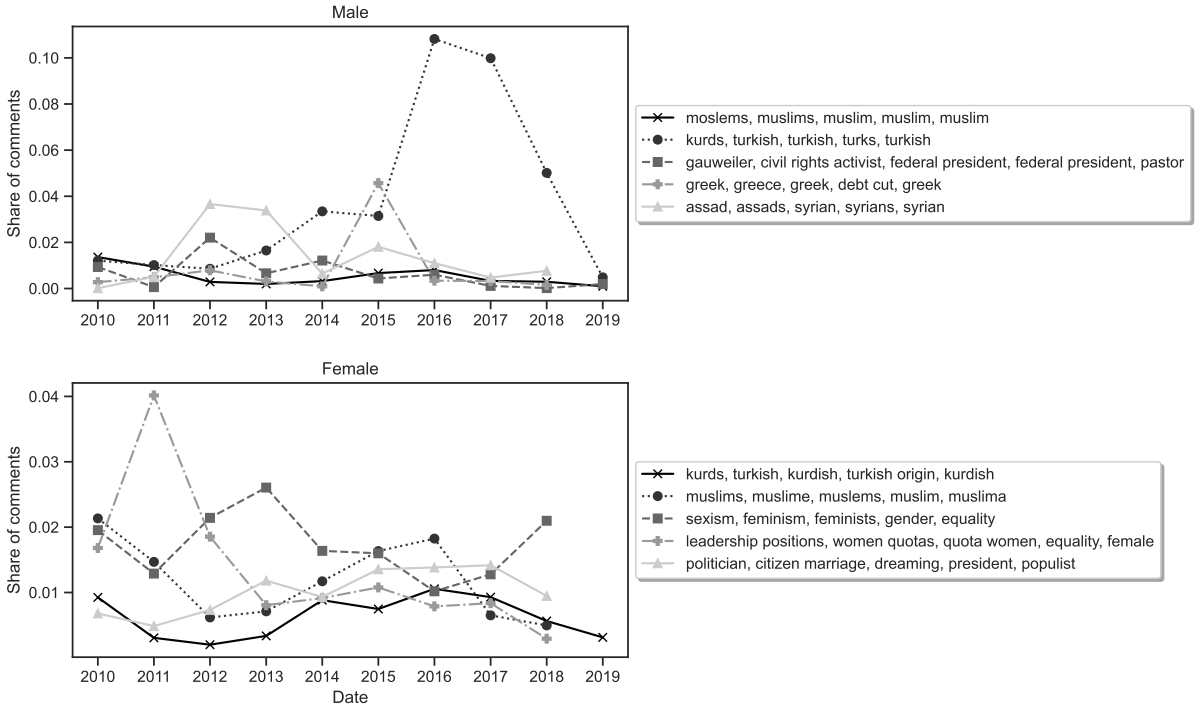


Figure 2.3: Development of the importance of the most frequently occurring topics over time, without comments on Angela Merkel.

2.3.2 Topic classification

To examine the prevalence and development of gender stereotypes in our data, we must assess whether and which (gender stereotypical) topics are being discussed in the comments, and whether the discussions center on women or men. This section illustrates our automated topic classification procedure, where we propose an innovative combination of dictionary and word embedding approaches to resolve crucial conceptual challenges such as gender bias, false positives, and false negatives. The gender (assessing whether comments discuss women or men), sentiment, and offensive language classifications of comments are specified in Sections 2.3.3 to 2.3.5 below.

2.3.2.1 Conceptual challenges

As argued above, the automated topic classification of comments could suffer from three pitfalls: gender bias, false positives, and false negatives. Gender bias is likely to arise in a naive supervised ML approach, where human coders manually classify whether comments cover a certain stereotypical topic or not. If this information was used to train a supervised ML algorithm, the algorithm would pick up existing gender stereotypes from the training data and transfer them to the sample of interest. E.g., suppose that comments in the training data discuss women more often than men in the context of *family*. A supervised ML algorithm would pick up this joint pattern and classify comments on women in the prediction sample accordingly. As a result, we would not be able to catch differences in gender stereotypes between comments in the training and in the prediction sample and, crucially, we would not be able to detect changes in gender stereotypes over time. In

other words, any topic classification procedure that is driven by the occurrence of gender in a specific comment is likely to yield biased results.

Dictionary methods that use curated lists of words or expressions related to a specific topic – typically put together by linguistic researchers – could solve this issue. Practitioners usually apply such methods by counting the occurrences of dictionary terms in a corpus (e.g., Tetlock, 2007). However, dictionary methods come with two disadvantages. First, they could yield false positives, as it is not trivial how to select or aggregate words in a corpus or a document to capture just the relevant and unambiguous ones. Second, they could yield false negatives, because the selection of words in a dictionary is naturally limited. In addition, dictionary methods are sensitive to prefixes, suffixes, typographical errors, or synonyms, especially when considering morphologically rich languages like German and error-prone online discussions.

In this paper, we propose a solution to these challenges by enriching unbiased dictionary methods with the flexibility and understanding of *word embeddings* (Mikolov et al., 2013). Word embeddings represent the semantic meaning of words by vectors in an n-dimensional space, where words with a similar meaning are represented by vectors that are close to each other.[11] Under the key assumption that words related to a specific (gender stereotypical) topic are clustered in the embedding vector space, this feature allows us to predict the topic(s) of a comment based on words that are semantically *similar* to those in an unbiased dictionary.

2.3.2.2 Procedure

Our topic classification procedure comprises two main parts – *training* and *prediction* – which consist of several smaller steps, respectively. Figure 2.4 provides an overview of the procedure, further details are discussed below.

Part 1: Training

Step 1: Dictionary pre-processing Part 1 of our topic classification procedure is based on *Linguistic Inquiry and Word Count Dictionaries* ("LIWC" henceforth), which provide extensive human-validated lists of words that correspond to certain topics.[12] E.g., the topic *work* includes words like *labor*, *office*, and *politician*, while the topic *family* includes words like *mother*, *brother*, and *childcare*. Following the recent literature (e.g., Fiske, 2010; Ellemers, 2018; Marjanovic et al., 2022), we identify six of the topics in LIWC as gender stereotypical: *work* and *money* for men, and *family*, *home*, *body* and *sexual* for women. Let T denote the set of all, and $T^{gender} \subset T$ the set of gender stereotypical topics in LIWC.

We start by removing all ambiguous words from all topics $t \in T^{gender}$. E.g., the topic *work* features words like *negotiate* and *request*, which could be related to workplace activities but also to other contexts. Thus, we let two Research Assistants independently

[11] See Gentzkow et al. (2019) and Ash and Hansen (2022) for intuitive discussions of word embeddings.

[12] Specifically, we use the German adaption *DE-LIWC2015* (Meier et al., 2019) of the English original developed by Pennebaker et al. (2015). For some supportive tasks, we also consider terms from the 2001 version of the LIWC (Wolf et al., 2008), which is based on the English original by Pennebaker et al. (2001).

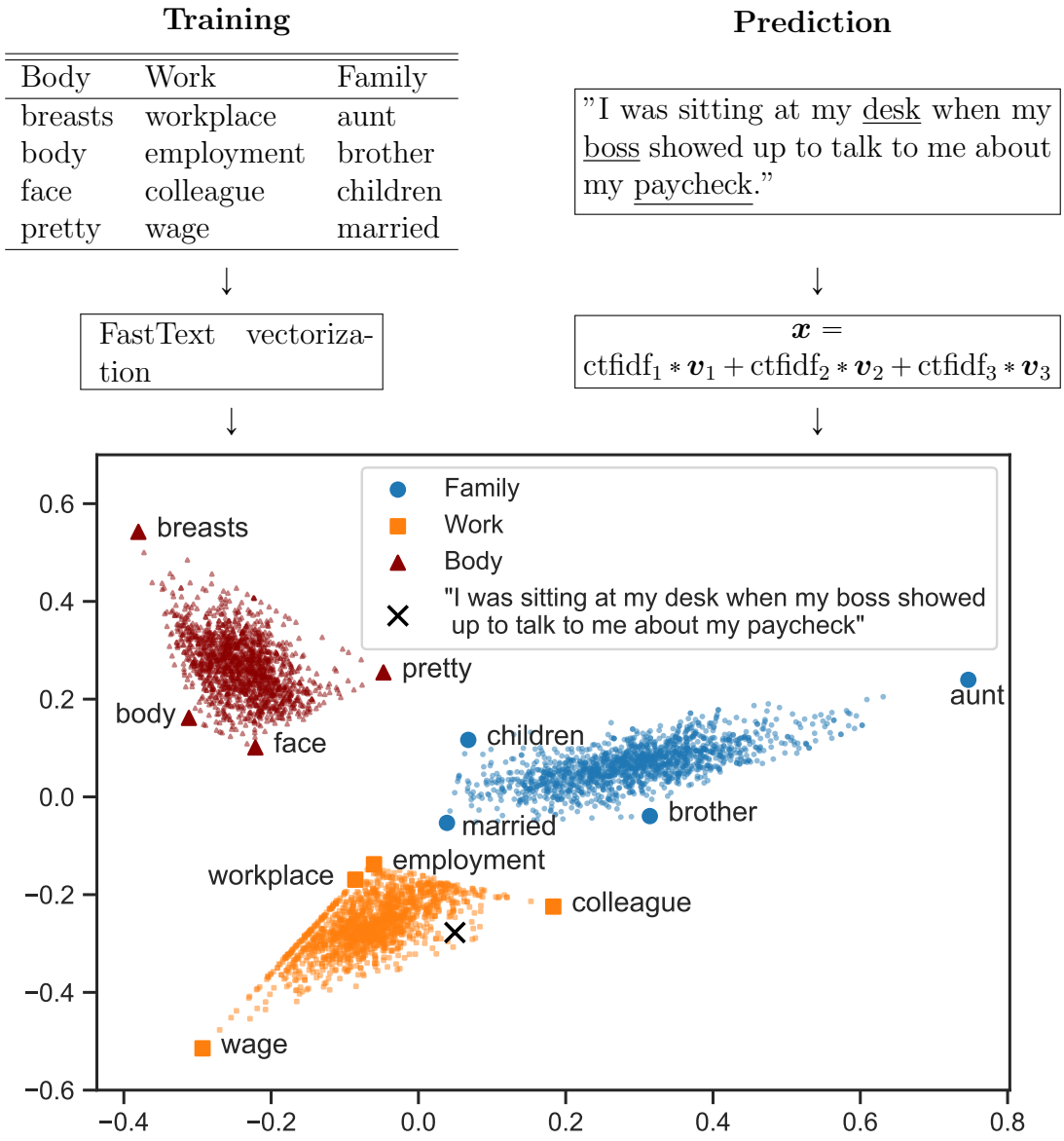


Figure 2.4: Stylized example of our topic classification procedure

Notes: In this example, we transform words from a dictionary featuring the topics *Body*, *Work*, and *Family* into their *FastText* word embedding representation. (In contrast to our actual model, we apply a *principal component analysis* to reduce the 300-dimensional vector space to just two dimensions; this enables us to visualize the data.) Large circles, squares, and triangles represent the word embeddings. Small circles, squares, and triangles represent linear combinations of the word embeddings, which we use as training data for a Support Vector Machine. Note that the word embeddings in this figure are based on our actual data.
The upper RHS displays a hypothetical comment with work-related content. Note, however, that none of its words appear in our stylized dictionary. We retrieve the comment's most important words with our *clustered tf-ifd* and compute their linear combination, using the normalized *clustered tf-ifd*-scores as weights. The corresponding vector is represented by the dark cross. Under the key assumption that words related to a specific topic are clustered in the embedding vector space, our trained model will predict the topic *work* with high, and the remaining topics with low probability.

decide which words unambiguously describe a certain topic and proceed only with those that both of them agreed upon. Further pre-processing steps include capitalization of nouns and replacing words that are designed to find match patterns with words that actually exist.[13]

Step 2: Word embeddings Next, we transform each of the remaining words from each topic $t \in T^{gender}$ into its real-valued 300-dimensional *FastText* word embedding (Bojanowski et al., 2017). Word embeddings are computed via neural network architectures that require huge amounts of data. Therefore, we do not compute the word embeddings ourselves, but rely on externally pre-trained data, as is standard in applied research (e.g., Garg et al., 2018; Kozlowski et al., 2019).[14] The *FastText* word embeddings are pre-trained by Mikolov et al. (2018), based on Common Crawl's web archive and the entire Wikipedia. This vast amount of training data ensures high quality results; moreover, since the ultimate goal of our procedure is to classify UGC, we perceive this kind of training data as particularly adequate.[15] The main difference between *FastText* and more traditional embedding methods like *Word2Vec* is that *FastText* is trained on n-grams rather than full words. This leads to practically no out-of-vocabulary words during prediction and performs well for morphologically rich languages like German.[16] Hence, *FastText* word embeddings are robust towards common dictionary concerns such as synonyms, compound words, prefixes, suffixes, and typographical errors.[17]

Step 3: Generate training data Based on the *FastText* word embeddings, we generate our training data. To this end, we split the word embeddings into three groups – training, test, and validation – where the training vectors are used as input for a supervised ML model. Crucially, we train a *separate* model for each topic $t \in T^{gender}$. Thus, each model will ultimately be able to predict whether a particular comment covers a particular topic t or not, but the individual topics are not mutually exclusive (e.g., a comment could be classified as being related to *work and* being related to *money*). Specifically, we conduct the following procedure *for each* of the six topics $t \in T^{gender}$:

Denote the focal topic as t^f (e.g., *work*). To generate *one* training observation i:

1. Randomly select one further topic $t^i \in T$, where t^i can be equal to the focal topic t^f or any other topic $t \neq t^f$ in T.[18]

[13]E.g., we replace *administrati** with *administration* and *analyse** with *analyse*.

[14]We use the *gensim* software library (Řehůřek and Sojka, 2010) to load and apply the pre-trained vectors.

[15]Note that the *FastText* word embeddings are likely to outperform any word embeddings that we train ourselves, simply because the training data used by Mikolov et al. (2018) is many times larger than our sample of comments.

[16]E.g., even if the term *Wahlumfrage* (election survey) does not occur in the training data, *FastText* is able to compute the embedding vector as a combination of its vectors for *Wahl* (election) and *Umfrage* (survey) and can thus capture similarities to both of its components.

[17]It has recently been argued that pre-trained word embeddings may be gender biased themselves (e.g., Gonen and Goldberg, 2019). While discarding this is beyond the scope of our paper, we believe that any potential gender bias in the word embeddings is smaller than the bias we would generate if we used a supervised ML approach on our data.

[18]Using the entire set of topics T instead of just T^{gender} enriches our collection of words from different

2. Pick three random word embeddings $\boldsymbol{v}^i$ from t^i and three random scalars w^i that add up to one. The linear combination of $\boldsymbol{v}^i$, using w^i as weights is given by

$$\boldsymbol{x}^i = w_1^i \boldsymbol{v}_1^i + w_2^i \boldsymbol{v}_2^i + w_3^i \boldsymbol{v}_3^i, \tag{2.1}$$

where $\boldsymbol{x}^i$ is a new vector in the same vector space and with the same dimensionality as the original word embeddings $\boldsymbol{v}^i$. Crucially, any $\boldsymbol{x}^i$ lies somewhere in between $\boldsymbol{v}^i$.

3. Finally, let y^i be a binary target variable, where $y^i = 1$ if $t^i = t^f$ (here: if t^i is equal to *work*), and $y^i = 0$ otherwise.

Steps [1] to [3] are repeated n times such that the y^i are roughly balanced with respect to being equal to 0 or 1.[19] Thus, we ultimately generate n training observations for each category $t \in T^{gender}$, where each training observation i consists of a 300-dimensional vector $\boldsymbol{x}^i$ (which, in turn, is a linear combination of the word embeddings $\boldsymbol{v}^i$) and a binary target variable y^i that indicates if $\boldsymbol{x}^i$ is a linear combination of word embeddings from the focal topic t^f or not.

Step 4: Training of the Support Vector Machine Next, we use the training observations from Step 3 as input for a supervised ML model (one model per topic $t \in T^{gender}$).[20] To this end, we let an ensemble of *Support Vector Machine* algorithms (SVM) use the $(n \times 300)$ input matrix $\boldsymbol{X}_t = (\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, ..., \boldsymbol{x}_t^n)^T$ to predict the vector of binary target variables $\boldsymbol{y}_t = (y_t^1, y_t^2, ..., y_t^n)^T$ for each $t \in T^{gender}$. Specifically, we consider an ensemble of three models for each t, where each of these is a sub-ensemble of SVM algorithms. Each algorithm in each sub-ensemble is trained on a different random draw of input data. We use 5-fold cross validation to tune the hyperparameters of all algorithms such that each algorithm within the sub-ensemble features an identical set of hyperparameters and differs only in the input data drawn at random. Then, we aggregate the three ensembles with the best performing sets of hyperparameters into a final ensemble. The SVM algorithms essentially search for borders that optimally separate observations belonging to t from observations that do not, slicing the vector space into areas that correspond to the individual topics $t \in T^{gender}$. Our key assumption here is that word embeddings from the same topic are clustered within the vector space.

Step 5: Intermediate Evaluation As an intermediate evaluation of our trained model, we come back to the yet unused validation word embeddings (see Step 3). In particular, we use our model to determine the probability with which each of these word embeddings corresponds to each topic $t \in T^{gender}$. Then, we compare our prediction with the actual topics that the validation word embeddings correspond to.[21]

contexts, whereby our algorithm will ultimately be better able to disambiguate them. We further support this approach with a short self-compiled list of words relating to *cars* and *politics*, since these topics play a dominant role in the UGC that we wish to classify.

[19] n is some multiple of the number of word embeddings in t^i. This topic specific multiplier value is found via hyperparameter tuning with 5-fold cross validation.

[20] Model training and prediction were executed with the Python library *scikit-learn* (Pedregosa et al., 2011).

[21] Recall that the actual topics of the validation word embeddings are known.

Table 2.1: Validation of the SVM ensemble

	Accuracy	Precision	Recall	F1
Work	0.947	0.815	0.791	0.803
Money	0.935	0.897	0.821	0.857
Family	0.988	0.885	0.885	0.885
Home	0.988	0.810	0.895	0.850
Body	0.946	0.899	0.860	0.879
Sexual	0.968	0.830	0.830	0.830

Notes: Prediction metrics for the validation word embeddings. *Accuracy* is the proportion of correct predictions. *Precision* is the proportion of correct positives. *Recall* measures the proportion of positives captured by the positive predictions. The f_1*-score* is the harmonic mean of precision and recall.

Table 2.1 displays four of the most frequently used evaluation metrics for binary classification. All of these metrics for all topics $t \in T^{gender}$ are close to 1, thus demonstrating that our trained model performs extremely well. Note, however, that the results in Table 2.1 are not (yet) informative about the topic classification of UGC, which we conduct in Part 2 of our procedure.

Part 2: Prediction

Step 1: Collapse comments by *clustered tf-idf* Before we can apply the trained model to our sample of interest, we must make the multi-word comments comparable to word-level embedding vectors. To this end, we use *tf-idf* (term frequency / inverse document frequency) to identify the most relevant words per comment. Specifically, since regular *tf-idf* ignores semantic similarity of words (which would reduce the flexibility and understanding that we gained through the word embeddings), we develop a *clustered tf-idf* approach, where words of similar meaning are considered together.

The *clustered tf-idf* comprises three steps. We start by computing regular *tf-idf* weights for all words in our corpus. Then, we use an unsupervised ML algorithm to cluster the words' *FastText* embedding vectors.[22] The algorithm is tuned to identify many clusters with few word embeddings, respectively, which assures that the embeddings within a cluster are semantically close to each other. We then aggregate the regular *tf-idf* weights of all words that correspond to the embedding vectors within one cluster to a *clustered tf-idf* weight, which, in turn, is assigned to all words within that cluster. If a cluster comprises just one word embedding, the *clustered tf-idf* corresponds to the regular *tf-idf* weight of the corresponding word.[23]

Based on the *clustered tf-idf* weights, we identify the three most relevant clustered word embeddings per comment.[24] Analogous to the words from LIWC, we transfer these nouns into their 300-dimensional *FastText* word embeddings. Then, we compute their

[22]More specifically, we use agglomerative hierarchical clustering (Murtagh and Contreras, 2012).

[23]See Appendix B.2 for further details.

[24]We identify nouns with the part-of-speech tagging capabilities of the *spacy* software library in Python. We focus on nouns, because they are less ambiguous in terms of their topic correspondence than adjectives or verbs.

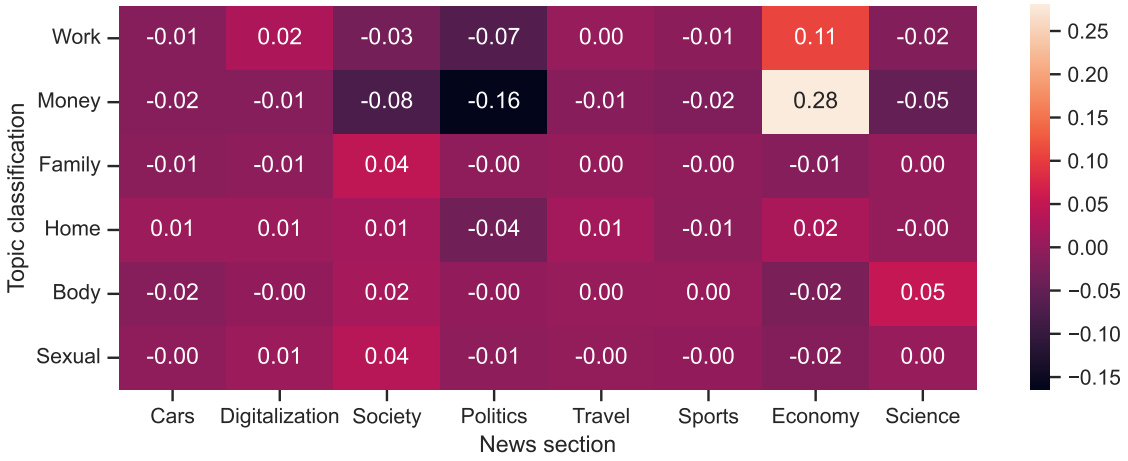


Figure 2.5: Correlation between news sections and topic classification

Notes: Values in cells are Pearson correlation coefficients between the binary classification predictions of comments and the binary indicator for the newspaper section the comment is located in. Brighter shadings indicate a larger positive correlation.

linear combination, using their normalized *clustered tf-idf* weights such that they add up to one. Thus, each comment is ultimately represented by a linear combination of word embeddings that is projected onto the same vector space as the training data, whereby we can apply the trained model from Part 1.

Step 2: Predict topics Finally, we use our trained model to predict the probability with which each of the collapsed comments discusses a specific gender stereotypical topic $t \in T^{gender}$. In particular, we classify a comment as discussing topic t if $\Pr(t) > 0.5$.[25] Note that the topics are *not* mutually exclusive; e.g., a comment could be classified as discussing both *work* and *money*. See Appendix B.1.1 for three examples of our topic classification.

2.3.2.3 Validation

We pursue two approaches to validate our automated topic classification. First, we consider the pairwise correlation between the predicted topics and the news outlet section that the comments were originally attached to (Figure 2.5). Plausibly, comments that we classify as covering the topics *work* and *money* are most strongly correlated to the news outlet's economy section, whereas comments that cover *family* and *sexual* appear most frequently in the news outlet's society, and comments on *body* in the science section.

Second, we apply our automated topic classification to chunks of text whose content is known. In particular, we screen the Wikipedia category tree[26] for the categories that best match the six gender stereotypical topics that we consider (Table B1 provides an overview).[27] E.g., Wikipedia articles from the category "working environment" are very

[25]Section 2.5 shows that our results do not hinge on this binary classification.
[26]See `https://en.wikipedia.org/wiki/Special:CategoryTree` (May 2022).
[27]We use the Wikipedia API to retrieve the first paragraph of all articles that belong to the selected

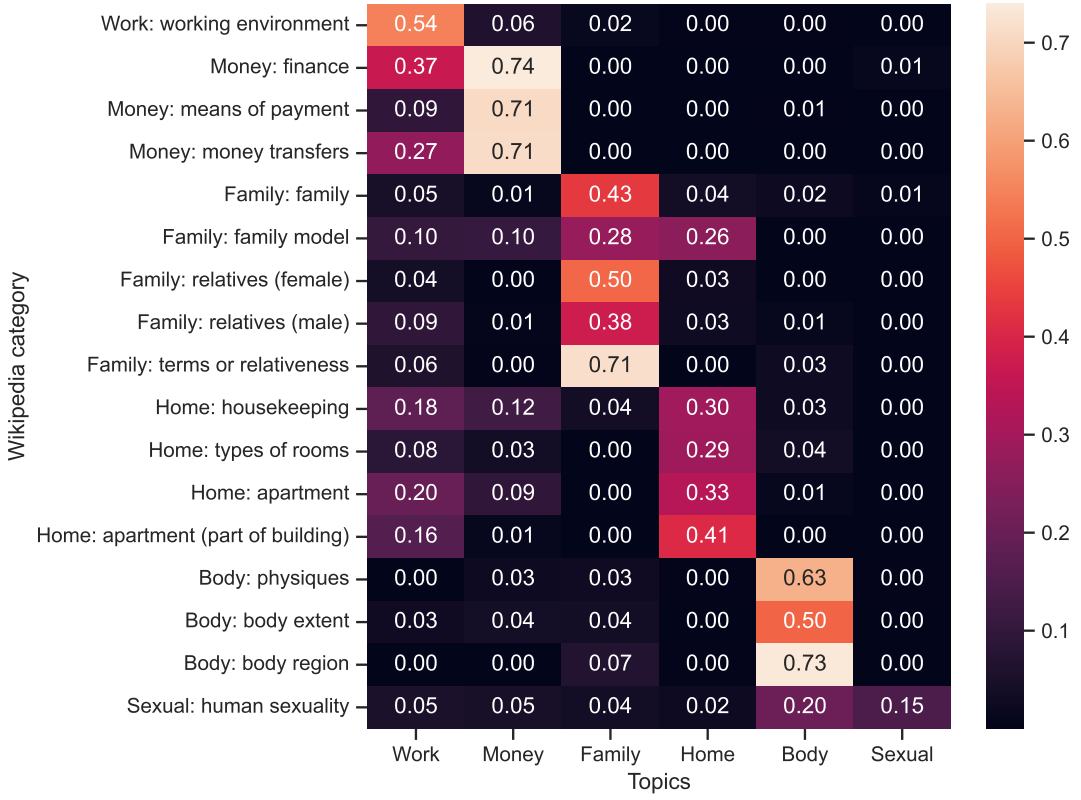


Figure 2.6: Topic classification of Wikipedia articles from known categories

Notes: The figure depicts the average predicted probabilities that articles from a specific Wikipedia category cover each of our gender stereotypical topics $t \in T^{gender}$. Brighter shadings indicate larger probabilities.

likely to cover the topic *work*; hence, our topic classification procedure should classify those articles accordingly.

Figure 2.6 shows that our procedure performs extremely well. In particular, we find that Wikipedia articles from categories that correspond to a certain topic $t \in T^{gender}$ are classified accordingly, while the average prediction probabilities for unrelated topics are low. E.g., our model predicts that Wikipedia articles from the categories "finance", "means of payment", and "money transfers" cover the topic *money* with a probability of up to 85%, and the remaining topics with a probability close to 0%.

2.3.3 Gender classification

In contrast to the more ambiguous (gender stereotypical) topics, the occurrence of men and women as part of the discussion in our comments is relatively explicit. E.g., if a comment mentions "Harry" or "Mr. Smith", it is clear that a man is being discussed. As a result, we can base the gender classification of our comments on a composite of simple dictionary approaches. To minimize the number of false positives, we restrict the

categories. From this list, we remove articles about individuals, interest groups, and redirects. Then, we apply our algorithm to classify each of the collected paragraphs.

procedure to few gender specific names and terms that are unambiguous in this context as well as to celebrities whose gender is publicly known. To minimize the number of false negatives, we combine three different dictionary approaches whose results complement each other.

2.3.3.1 Procedure

The gender classification of our comments is based on three dictionary approaches:

First names We start by retrieving a list of the 100 most popular German male and female first names from *The Society for the German Language*'s website and remove ambiguous names such as *Ernst* ("serious").[28] Then, we search each comment for the occurrence of one or several male or female first names. If a comment features at least one female first name, it is classified as *female*, if it features at least one male first name, it is classified as *male*, and if it features no popular first name at all, it is classified as *none*. Note that a comment could be classified as both *male and female* at this stage; ties are resolved when we compile the results from all three approaches.

Gender specific terms Second, we search the comments for unambiguous gender specific terms like *lady* or *gentleman*.[29] Analogous to the above procedure, we classify a comment as *female* (*male*) if it contains at least one of these terms, and as *none* otherwise.

Celebrities Third, we let a Research Assistant read several thousand comments and compile a list of all celebrities that she came across (e.g., Donald Trump, Angela Merkel, Beyoncé). Based on this list, we searched all comments for the occurrence of celebrities whose gender is publicly known.[30] As above, we classify a comment as *female* (*male*) if it features at least one female (male) celebrity, and as *none* otherwise.

Composition We compile the results from our dictionary approaches in three steps. We first consider consonant classifications. In particular, we ultimately classify a comment as *female* (*male*) if at least one of the dictionary approaches classifies the comment accordingly, and the other approaches either agree or classify the comment as *none*.

In a second step, we resolve conflicting classifications *across* our dictionary approaches (i.e., if one approach classifies a comment as *male* and another approach as *female*). Since gender specific terms are less ambiguous than first names, and celebrities are less ambiguous than gender specific terms, our third approach overrules the second one, and the second approach overrules the first. Section 2.5 demonstrates that our results are robust to alternative composition rules, such as the first names or the gender specific terms overruling the other approaches.

[28] *Gesellschaft für deutsche Sprache e.V.*, see `https://gfds.de/vornamen/beliebteste-vornamen/#` (Nov 2022) for further details.

[29] In particular, we search for the gender specific male terms *herr, mann, männ* and the gender specific female terms *frau, dame, weib, mädchen, fräulein*.

[30] The procedure yields a total of 1,491 male and of 511 female celebrities. When we search the comments , we take different spellings and spelling mistakes of the celebrities into account.

Finally, we resolve ties *within* our composite classification (i.e., if a comment is classified as both *male and female* after resolving conflicting classifications across the approaches). Specifically, if we find that a comment discusses both men *and* women, we set its classification to *none*. Section 2.5 shows that our results are robust to classifying such cases according to the majority of male/female first names, gender specific terms, and celebrity occurrences (i.e., classify a comment as *female* if there are more female than male classifiers and vice versa). See Appendix B.1.1 for three examples of our gender classification.

2.3.3.2 Validation

As argued above, the explicit discussion of men and women in our comments joint with the careful selection of terms for our dictionary approaches curtails the risk of generating false positives and negatives. To validate the performance of our gender classification procedure nonetheless, we use a Lasso-Logistic propensity score model and examine the words that are most predictive for comments classified as *male* or *female*. Specifically, we draw a random sample of $3,000$ *male, female,* and *none* comments, respectively, and use the trained *tf-ifd* from Section 2.3.2 to vectorize the comments.[31] Then, we run two separate Lasso-Logistic regressions, where we use the *male* classifier as dependent variable in the first, and the *female* classifier as dependent variable in the second regression.

Table 2.2 displays the ten most predictive terms for comments classified as *female* (column 1) and *male* (column 2), respectively. The results are compelling: while words such as *wife*, *mother*, *family*, and *child* are most predictive for comments classified as *female*, words like *money*, *war*, and *president* are most predictive for comments classified as *male*. This does not only validate our gender classification, but also prefigures our main results on gender stereotypes that we present in Section 2.4.

2.3.4 Sentiment

To examine if gender stereotypes are driven by hostile or benevolent sexism (Glick and Fiske, 2001, 2018), we also determine the sentiment of our comments. To this end, we apply Latent Semantic Scaling (LSX, Watanabe, 2021) to compute a sentiment score for each comment. Similar to our topic classification, LSX adopts a polarity lexicon, where seed words are assigned to a positive or negative class. These seed words are then transferred to their word embedding representation, and the polarity of other words can be inferred from the similarity of their word embeddings to the embeddings from the dictionary.

To apply LSX to our analysis, we use the SentiWS sentiment dictionary (Remus et al., 2010), which provides an extensive list of German words along with a polarity score ranging from -1 to 1. We restrict the analysis to words with an absolute score above 0.5, which gives us about 120 words, and use *FastText* to transfer these words into their word embeddings. Then, we compute the similarity of all nouns, verbs, adjectives, and adverbs in each of our comments with the word embeddings from the dictionary, weight the words with the corresponding polarity scores, and use the *clustered tf-idf*

[31] As in our main specification (see Section 2.4), we exclude all comments about Angela Merkel from the analysis.

Table 2.2: Most predictive words for *female* and *male* comments

(1) *female*	(2) *male*
wife	money
family	chancellor
society	war
child	party
life	politics
mother	politician
victim	president
party	law
quota	government
law	people

Notes: This table displays the ten most predictive words for comments classified as *male* or *female*. The words are obtained via two separate Lasso-Logistic propensity score models based on a random sample of 9,000 comments, where 3,000 are classified as *male*, 3,000 are classified as *female*, and 3,000 are classified as *none*.

method from above to compute an aggregate sentiment score for each of our comments. Finally, we standardize the comment-level sentiment scores such that comments that are more negative than the average feature a negative, and comments that are more positive than the average feature a positive score.

2.3.5 Offensive language

Since we are mainly interested in subtle forms of gender stereotypes, we classify all comments that use offensive language to distinguish them from more common speech in our subsequent analyses. To this end, we employ a multilingual *BERT* model (Devlin et al., 2018), i.e., a large pre-trained language model that we fine-tune for the supervised prediction of offensive language in our comments. To this end, we use German Tweets from Wiegand et al. (2018) and Struß et al. (2019), which come with a crowdsourced indicator for offensive language as training data.[32] Then, we apply the trained model to our sample of comments to predict the probability with which each comments features offensive language. Analogous to our topic classification, we classify a comment as featuring offensive language if $\Pr(offensive) > 0.5$.

While we cannot validate the performance this model on our comments, we can validate how well it performs on a sample of held out validation Tweets and assume that the Tweets and the offensive language in them are sufficiently similar to our comments. Table 2.3 shows that the model produces relatively few false positives but does miss out on some offensive posts.

[32]The website for this labelling task defines offensive language as "hurtful, derogatory or obscene comments made by one person to another person" (https://fz.h-da.de/iggsa).

Table 2.3: Evaluation metrics offensive comments

Accuracy	Precision	Recall	F1
0.80	0.74	0.60	0.66

2.4 Results

This section presents the results from applying the topic, gender, sentiment, and offensive language classification to our sample of interest. We start by providing (static) descriptive evidence, then we present the results on the prevalence and development of gender stereotypes over time.

2.4.1 Descriptives

Gender classification Since our main analysis is based on comments that are classified as either *male* or *female*, we start by considering the results of our gender classification. From our initial sample of $7,345,166$ comments, $1,375,252$ are classified as discussing either women or men. From these, we exclude all comments that mention Angela Merkel, as she is likely to be an outlier in terms of the subtle and unconscious gender stereotypes that we wish to examine (see also Figure 2.2).[33] This reduces the number of comments for our main analysis to $1,162,735$, where $200,261$ comments (17.22%) are classified as *female*, and $962,474$ (82.78%) are classified as *male*.

Topic classification Based on the $1,162,735$ comments from above, Table 2.4 summarizes the results of our topic classification. The topics that appear most frequently in our main sample are *work* and *money*, i.e., those that we perceive as stereotypical male. In contrast to that, topics that we identify as stereotypical female – *family*, *home*, *body*, and *sexual* – appear relatively seldom. To further check the validity of our main results, we also introduce two placebo topics – *time* and *space* – which are arguably unrelated to gender. Hence, when we compare how often men and women are mentioned in the context of *time* and *space*, we should not be able to observe any differences between these groups.

Some of our gender stereotypical topics are conceptually similar (e.g., *family* and *home*). To take this into account – and to present the prevalence and development of gender stereotypes as concisely as possible – we pool comments that are classified as *work* or *money* (or both) as *professional*. Analogously, we pool *family* and *home* as *domestic*, *body* and *sexual* as *physical*, and *time* and *space* as *placebo*. Section 2.5 shows that our results are qualitatively similar when we consider each of those topics individually.

Figure 2.7 displays the proportion of comments classified as *professional*, *domestic*, *physical*, and *placebo* for *male* and *female* comments, respectively. While the proportion of *female* comments classified as *professional* is smaller than for *male* comments, it is considerably larger for comments classified as *domestic* and *physical*, which strongly suggests that gender stereotypes exist in our data. The difference between *male* and *female* comments for *placebo*, in contrast, is negligible.

[33] We use a simple dictionary approach to identify referrals to Angela Merkel. Section 2.5 provides a robustness check, where we keep comments on her in our sample.

Table 2.4: Topic classification

Topic	No. comments	Share
Original		
work	113,334	9.75%
money	157,618	13.56%
family	19,145	1.65%
home	15,141	1.30%
body	67,965	5.85%
sexual	5,096	0.44%
time	2,916	0.25 %
space	18,180	1.56%
Pooled		
professional	255,690	21.99%
domestic	33,932	2.92%
physical	72,714	6.25%
placebo	21,087	1.81%

Notes: Results of our topic classification. Note that the topic classification is not mutually exclusive, i.e., a comment could be classified as covering zero, one, or several topics.

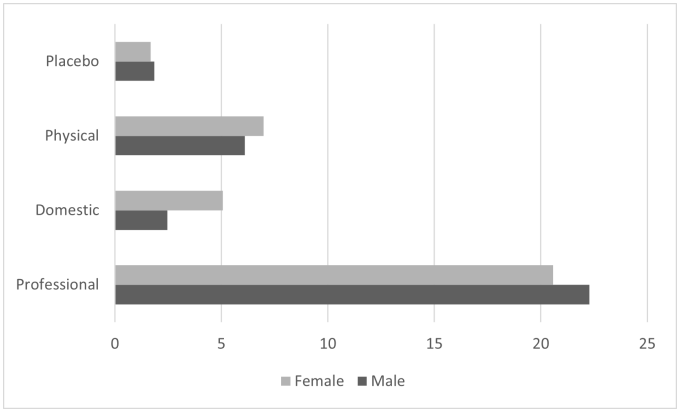


Figure 2.7: Pooled topic classification by gender (in %).

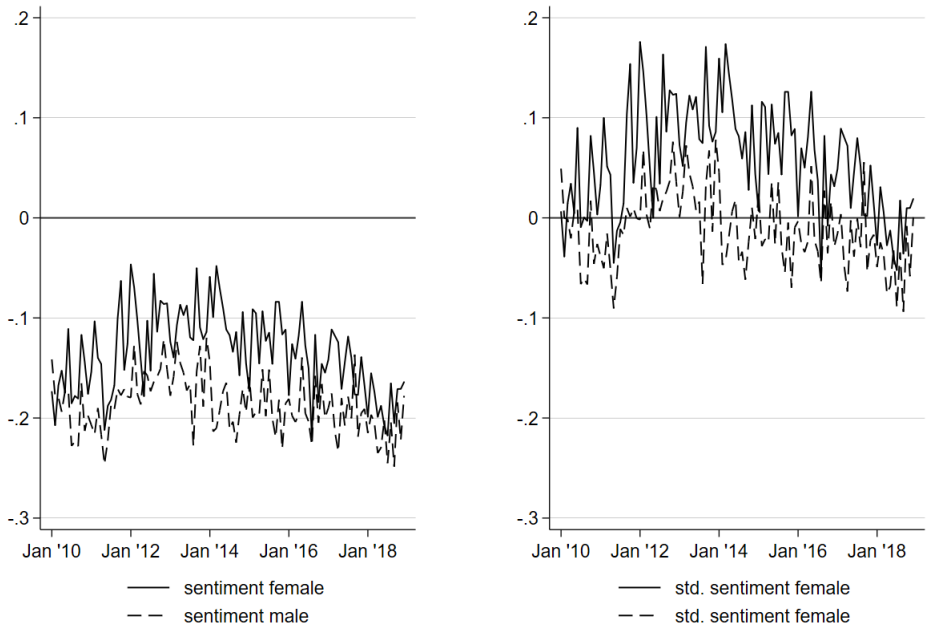


Figure 2.8: Average sentiment scores per month and gender. Left panel: raw sentiment scores. Right panel: standardized sentiment scores.

Sentiment Figure 2.8 shows the results of our sentiment classification. The left panel depicts the average sentiment score for comments classified as *male* or *female* in each month of our observation period. We find that both types of comments have negative sentiment scores on average, where comments classified as *male* are usually more negative than comments classified as *female*. The development of sentiment is mostly parallel for *male* and *female* comments: the average sentiment scores increase until about Jan 2014, then decline steadily with a particularly sharp drop for *female* comments by the end of 2017.

To facilitate the interpretation of our sentiment score, we standardize the values to have a mean of zero and a standard deviation of one. The right panel in Figure 2.8 shows the results: after standardization, the average sentiment score for comments classified as *male* is close to and fluctuates around zero, whereas the average sentiment score for comments classified as *female* is largely positive. As illustrated above, we use these standardized sentiment scores as weights for our comments. In particular, we multiply each comment i that is classified as covering a gender stereotypical topic $t \in T^{Gender}$ with $(1 + \text{std\_sentiment\_score}_i)$. Thus, comments with average sentiment are given the same weight as in our main analysis, whereas more benevolent comments are given larger, and more hostile comments are given lower weight than before.

Figure 2.9 displays the mean standardized sentiment scores by (gender stereotypical) topic and gender. Throughout all topics, we find that comments about women are on average more positive than comments about men, where the absolute differences are largest for comments that discuss domestic issues or physical appearance.

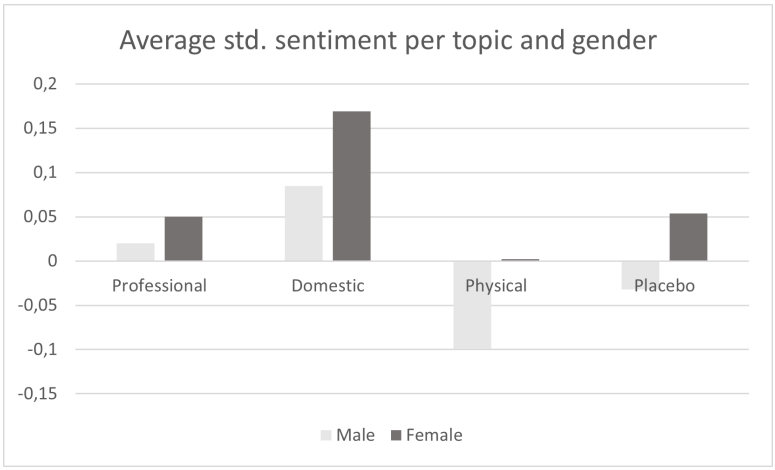


Figure 2.9: Average standardized sentiment scores per topic and gender.

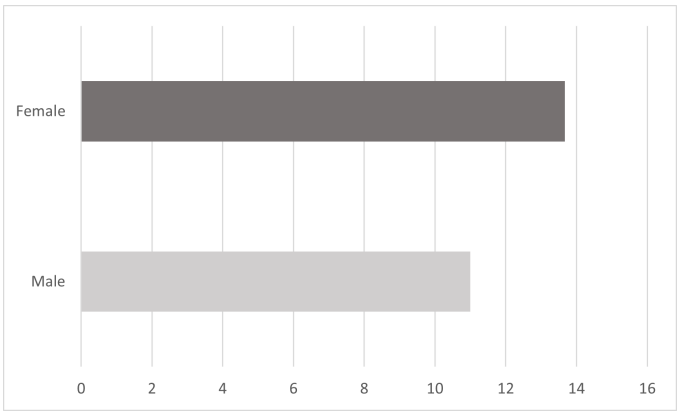


Figure 2.10: Percentage of offensive comments by gender

Offensive language We find that 133, 266 or (11.46%) of our comments are classified as *offensive*. In particular, 13, 67% of comments that we classify as *female*, and 11% of comments that we classify as *male* feature offensive language (Figure 2.10). Note that this result does not conflict with our findings on sentiment: in particular, although there are relatively more offensive comments on women, their sentiment is on average more positive than the average sentiment of offensive comments about men.

2.4.2 Prevalence and development of gender stereotypes

2.4.2.1 Index

One core contribution of our paper is to document the prevalence and development of gender stereotypes over the course of almost a decade. To this end, we compute an index that captures the degree to which gender stereotypes exist in our data at a given point in time. More specifically, we consider each of our pooled (gender stereotypical) topics t in a particular month τ. For that topic and month, we count how many comments i are classified as *female*, and how many are classified as *male*. To take into account that there are generally fewer comments about women than about men, we normalize these counts with the absolute number of *female* and *male* comments in month τ, respectively.

Finally, we compute the difference between these normalized counts for each topic t and month τ:

$$index_{t,\tau} = \frac{\sum_i (female_{i,\tau} \cap t_{i,\tau})}{\sum_i female_{i,\tau}} - \frac{\sum_i (male_{i,\tau} \cap t_{i,\tau})}{\sum_i male_{i,\tau}}. \tag{2.2}$$

If women are mentioned relatively more often than men in the context of a specific topic t in month τ, the index in Equation (2.2) is positive. If, in contrast, men are mentioned relatively more often than women, the index in Equation (2.2) is negative. In other words, a negative index for the topic *professional*, as well as positive indices for the topics *domestic* and *physical*, would be in line with the existence of gender stereotypes in our data.

2.4.2.2 Main results

Baseline Figure 2.11 shows our main results, which document the prevalence and persistence of gender stereotypes in UGC. We find that men are discussed more often in the context of *professional* topics than women (index predominantly negative), and that women are discussed more often in the context of *domestic* and *physical* topics than men (index consistently positive). This prevalence of gender stereotypes is relatively stable over time. In particular, our indices remain roughly within the same range over the entire observation period of nine years. However, while we observe no time trend for our index on *domestic*, gender stereotypes in the context of *professional* and *physical* seem to diminish slightly. Specifically, the index for *professional* moves closer towards zero and is even temporarily positive after Jan '13. The index for *physical* approaches zero by the end of 2017. Reassuringly, the index for our placebo topics is close to zero over the entire time period.

The short-term development of our indices can in parts be linked to eminent national and international events. E.g., the more gender balanced discussion on *professional* topics in 2013/14 coincides with the famous National Socialist Undergrounds (NSU) Trial that centered on the alleged (female) terrorist Beate Zschäpe and gained huge media attention in Germany. Similarly, the downward movement for our index on *physical* by the end of 2017 coincides with the global #MeToo-movement, and the 2018 instances where it becomes negative could also be explained with the football world cup. In sum, however, our indices remain relatively stable over time, suggesting that gender stereotypes prevail irrespective of what is happening around the world.

Note that our indices capture the ultimate prevalence of gender stereotypes at any given point in time, but they remain agnostic about what drives the differences between women and men. E.g., we show that women are discussed relatively less often in the context of professional issues than men, and that this difference diminishes over time, but the index as such is not informative about whether this trend is caused by specific events, more prominent female figures in the public debate, a change in users' perception of gender roles over time, or the exit/entry of users with more or less gender stereotypical perceptions, to name just a few potential explanations. We consider this as a feature, rather than a short-coming, of our main analysis. In particular, our main objective is to provide a clean documentation of the prevalence and development of gender stereotypes over time, which is just equivalent to studying the aggregate effect of all potential mechanisms mentioned above. In other words, if our main interest is to measure the

Figure 2.11: Main results

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The blue dotted line corresponds to the index as illustrated in Section 2.4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

absolute prevalence of gender stereotypes in UGC – which is arguably what matters most for public policy – rather than studying selected aspects of it, potential mechanisms that could drive the index play an interesting but secondary role. We further discuss this issue in Section 2.4.2.3 below, where we conduct regression analyses that control for comment and user characteristics.

Sentiment Figure B3 presents our sentiment-weighted indices, which are very much comparable to our main results. While the indices for *professional* and *physical* are slightly more positive than before, the index for *domestic* is largely unaffected. This is in line with what we report in Figure 2.8 and the results on offensive language that we discuss below. Hence, there is just small (if any) evidence for the presence of benevolent, and no evidence for hostile sexism in our data.

Offensive language Figure B4 shows that our indices are as good as unaffected when we exclude comments classified as *offensive* from our data, emphasizing again that our approach captures subtle and unconscious gender stereotypes that are not expressed in terms of explicit harassment. In addition, Figure B4 illustrates that our main results are robust to potential time variation in forum moderation policies: even when we remove

every comment that features offensive language, our main results prevail.[34]

News articles Our main argument for studying the prevalence and development of gender stereotypes in UGC is that users' anonymity allows them to voice what they think but would otherwise not say. However, it could be that the comments merely take up gender stereotypes from the news articles that they were originally attached to. In this case, the above results would not be informative about subtle and unconscious stereotypes on behalf of the users.

To demonstrate that the prevalence and development of gender stereotypes in our comments is independent of the news articles, we retrieved the text body of all articles that the comments are attached to.[35] Then, we classify the articles analogous to the procedures that we describe in Section 2.3.2.2 and compute indices as specified in Section 2.4.2.1.

Figures B5 shows that although the indices for UGC and news articles move to some extent in parallel, the latter fluctuate more around zero, indicating that news coverage is gender balanced. Hence, while it is plausible that the two types of indices have a similar shape – since both are likely to be affected by the same eminent events around the world – we find no evidence for gender stereotypes in the news articles, and hence conclude that the users' discussion reflects their own inherent, and not just potential gender stereotypes from the news articles.

2.4.2.3 Regression analyses

To further explore the prevalence and development of gender stereotypes over time, this section provides the results from two types of regression analyses, where we control for comment and user characteristics. This allows us to study if and to what extent our main results are driven by observable features such as length of the comment, news outlet section, or user fixed effects.

We start by estimating the regression equation

$$Topic_{i,\tau} = \beta_0 + \beta_1 female_i + \beta_2 month_\tau + \beta_3 female_i * month_\tau + \theta X_i + \lambda_s + \lambda_u + \varepsilon_{i,\tau} \quad (2.3)$$

by OLS, where $Topic_{i,\tau}$ indicates whether comment i is classified as covering one of our pooled gender stereotypical topics, respectively, $female_i$ is a dummy equal to one if comment i is classified as *female*, and $month_\tau$ is a continuous variable capturing a linear time trend. The vector X_i comprises length of a comment, sentiment, and an indicator for offensive language. Finally, λ_s and λ_u capture news section and user fixed effects. Standard errors are clustered on the thread level. The parameters of interest are β_1 and β_3. Specifically, β_1 measures the average difference in the propensity to be mentioned in the context of a particular gender stereotypical topic between women and men for the entire observation period of almost a decade, conditional on our controls. β_3, on the other hand, measures if this difference has risen or fallen over time.

Table 2.5 shows the OLS estimates using each of our pooled gender stereotypical topics as dependent variable. Consistent with the main results in Section 2.4.2.2, the estimate

[34] Our main results are also robust to applying even stricter classifications of offensive language.

[35] $N = 70,235$, equivalent to the number of threads that we consider in our main analysis.

for $female_i$ is negative for *professional*, positive for *domestic* and *physical*, and close to zero for the placebo topics, irrespective of the empirical specification. Similarly, we find evidence that the prevalence of gender stereotypes in the context of professional topics declines over time: the corresponding estimate is positive and statistically significant. In contrast to that, the evidence for a decline in gender stereotypes in the context of domestic or physical issues is less clear. Although the corresponding estimates are negative and statistically significant (also owing to our large sample size), they are extremely small both in absolute terms and also relative to our estimates for $female_i$. Interestingly, the estimates hardly change when we include our fixed effects, indicating that the results are driven by variation within users and news sections. In other words, it is not the case that users or news sections with smaller gender bias become more important over time, but that the same users and news sections undergo (small) changes.

Table 2.5: Regression results

	Professional			Domestic		
	(1)	(2)	(3)	(4)	(5)	(6)
$female_i$	-0.0301***	-0.0243***	-0.0228***	0.0291***	0.0262***	0.0234***
	(0.00186)	(0.00178)	(0.00189)	(0.00097)	(0.00097)	(0.00101)
$female_i * month_\tau$	0.00030***	0.00031***	0.00025***	-0.00008***	-0.00007***	-0.00006***
	(0.00003)	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00002)
X_i		Yes	Yes		Yes	Yes
λ_s		Yes	Yes		Yes	Yes
λ_r			Yes			Yes
N	1,148,313	1,148,313	1,094,588	1,148,313	1,148,313	1,094,588
	Physical			Placebo		
	(7)	(8)	(9)	(10)	(11)	(12)
$female_i$	0.0129***	0.00989***	0.00908***	-0.00090	-0.00050	-0.00049
	(0.00117)	(0.00116)	(0.00123)	(0.00057)	(0.00057)	(0.00061)
$female_i * month_\tau$	-0.00010***	-0.00009***	-0.00009***	-0.00001	-0.00002	-0.00001
	(0.00002)	(0.00002)	(0.00002)	(0.00001)	(0.00001)	(0.00001)
X_i		Yes	Yes		Yes	Yes
λ_s		Yes	Yes		Yes	Yes
λ_r			Yes			Yes
N	1,148,313	1,148,313	1,094,588	1,148,313	1,148,313	1,094,588

Notes: Robust standard errors in parentheses. Standard errors are clustered on the thread level. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

In a second regression analysis, we regress $Topic_i$ on X_i, λ_s, and λ_u alone and replace the topic indicator $\cap t_{i,\tau}$ in the computation of our gender stereotype index from equation (2.2) with the residuals $\hat{\varepsilon}_{i,\tau}$ from that regression. The idea is that these residuals represent the probability of a specific gender stereotypical topic conditional on observed comment characteristics and user and news section fixed effects.[36] In line with the results from Table 2.5, Figure B6 shows that our indices are closer to zero but qualitatively similar to those that we present in Section 2.4.2.2, suggesting that the prevalence and development of gender stereotypes in UGC is not predominantly driven by any of our controls.

[36]Section 2.5 demonstrates that it hardly makes a difference whether we base our indices on topic indicators or (continuous) predicted probabilities.

2.5 Robustness checks

Angela Merkel Our main analysis excludes all comments on Angela Merkel as outliers. Figure B7 shows that our results are qualitatively comparable when we keep those observations in our sample. In particular, our index for *professional* remains predominantly below, and our index for *domestic* predominantly above zero. In contrast to our main results, the index for *physical* is *negative*; moreover, the index for *domestic* and is closer to zero than above. This finding is intuitive: Angela Merkel does not correspond to classic female gender stereotypes and is seldom discussed in the context of family, home, and physical appearance. Thus, considering comments on her in the analysis shifts these indices downwards.

Interpreting the index for *professional* requires closer examination. From Jan '10 to about Jan '15, the index is on average closer to zero than in Figure 2.11, i.e., the discussion is more gender balanced. Afterwards, the index is on average further away from zero than in Figure 2.11, i.e., the discussion becomes less gender balanced. A plausible explanation is Merkel's prominent role in the refugee crisis starting in Spring 2015. In particular, Merkel pursued a very warm and welcoming policy towards Syrian refugees and thereby triggered plentiful debates among politicians and the public, including users from our discussion forum. As a result, Merkel appeared in many comments that are not related to work or money, thus shifting the index further away from gender balance.

Alternative topic classifications Next, we explore the robustness of our results to alternative topic classifications. In particular, we show that we obtain similar results when we consider each gender stereotypical topic separately (i.e., when we do not pool related topics), and when we use a non-binary topic classification.

Figure B8 displays our index from Section 2.4.2.1 for each gender stereotypical topic $t \in T^{gender}$ as well as for our two placebo topics *time* and *space*. With the exception of *work*, all indices are similar to those that we present in Section 2.4.2.2. Specifically, the index for *money* is predominantly negative, the indices for *home*, *family*, *body*, and *sexual* are predominantly positive, and the indices for *time* and *space* are close to zero. In contrast to our main results, the index for *work* fluctuates around zero, indicating that gender stereotypes in the context of *professional* are mainly driven by gender stereotypes in discussions about money-related issues.

The indices in Figure B9 are based on a non-binary topic classification. Specifically, we do not assign a dummy equal to one if our algorithm predicts that comment i covers topic t with $Pr(t) > 0.5$, but use the predicted probabilities $Pr(t)$ themselves to compute the index from Section 2.4.2.1. This makes the indices harder to interpret, but also preserves information that would otherwise get lost (e.g., if the predicted probabilities for a certain topic are often positive, but smaller than 0.5).

Figure B9 shows that, with the exception of *work*, our indices are nearly unaffected. In particular, the predicted probabilities $Pr(t)$ are either close to zero or close to one, whereby using them instead of dummies does not make much of a difference. In contrast to that, comments classified as *female* often feature a small but positive probability to cover work-related issues. In consequence, the index for *work* is consistently above zero, suggesting that women are *more* likely to be discussed in the context of work than men. We perceive this result as slightly misleading, though. In particular, small but positive

predicted probabilities to cover a specific topic are more indicative of a comment *not* covering than actually covering that topic and should be interpreted accordingly (which is, e.g., facilitated by a binary classification as in our main specification).

Alternative gender classification We present the results from three alternative gender classification procedures. First, we re-consider ties *within* our composite gender classification. Specifically, we do not exclude observations that are classified as both *male* and *female* from the analysis, but resolve the ties with respect to the number of male and female instances within one comment. In particular, we count the absolute number of male and female first names, gender specific terms, and celebrities, and classify a comment as *male* if the former overweights the latter and vice versa. Only if the absolute number of male and female instances is exactly equal to each other, the comment is classified as *none* and dropped from the sample. Figure B10 shows that our results are as good as unchanged when we base our indices on this alternative gender classification.

Second and third, we re-consider ties *across* our composite gender classification. In particular, we let (i) gender specific terms and (ii) first names overrule the results from the other approaches. As above, our main indices remain nearly unchanged with this new specification, and are thus omitted.

2.6 Conclusion

Gender stereotypes – i.e., general expectations about attributes, characteristics, and roles of women and men – pose an important hurdle on the way to gender equality. It is difficult to quantify the problem, though, since gender stereotypes are not always conscious, and even if they are, they may not be openly expressed. This paper exploits the anonymity of UGC to overcome such challenges. In particular, we develop an innovative ML-based procedure that enriches unbiased dictionaries with the flexibility and understanding of word embeddings to classify more than a million user-written comments from a major German discussion forum in terms of (stereotypical) topics, gender, and sentiment. Based on that, we can document the prevalence and development of gender stereotypes over time.

We find strong evidence for the existence and persistence of gender stereotypes in our data. Specifically, we show that men are discussed relatively more often in the context of work and money than women, while women are discussed relatively more often in the context of family, home, and physical appearance than men. While the prevalence of gender stereotypes associated to male topics like work and money diminish slightly, gender stereotypes associated to female topics such as family and home persist over time. This result is supported by regression analyses that control for comment characteristics as well as for user and news section fixed effects. The results are also robust to excluding offensive language from our data, and they are not driven by potential stereotypes in the news articles that the comments were originally attached to. Moreover, we find just small evidence for benevolent, and no evidence for hostile sexism as drivers of gender stereotypes.

Assessing the prevalence and development of gender stereotypes in our society is a necessary requirement to take further actions towards gender equality. In particular, it is

important to understand more subtle and unconscious stereotypes, as these are harder to address than explicit discrimination and harassment. At the same time, however, subtle gender stereotypes are way more difficult to measure. As far as we know, our paper is the first that leverages the anonymity of UCG for a clean and extensive analysis of the prevalence and development of (subtle and potentially unconscious) gender stereotypes over time. We thus advance a paramount societal debate concerning academics, policy makers, and the general public. Our paper presents sharp evidence for the existence of gender stereotypes in UGC. Above all, however, our findings indicate that gender stereotypes prevail despite all measures that have been taken so far and despite global social media movements like #MeToo, thus calling for intensified efforts or alternative remedies.

We develop a novel procedure for the topic classification of UGC that can be applied far beyond this paper. E.g., our procedure allows for topic classification in the absence of labeled training data and for flexible dictionary classification even with small dictionaries. These features are particularly useful in the context of novel and unconventional data such as text from social media and other online platforms, languages where extensive dictionaries do not exist, and all types of text as data that have rarely been studied before and thus do not exhibit large training data. The procedure is especially useful in contexts where classic supervised ML models could learn certain patterns from the training data and transfer them to the sample of interest, which is problematic if changes in such patterns are the of interest by themselves. To further support research in that direction, our method is available as a Python package on `https://github.com/VFMR/WEELex`

Our paper has several limitations that open up avenues for further research. First, while we document the prevalence and development of gender stereotypes in UGC, we stay agnostic about their relation to actual attributes of women and men. In other words, assessing whether and to what extent gender stereotypes in UGC are a precise or biased reflection of real world circumstances is beyond the scope of our paper. However, gender stereotypes in terms of people's *expectations* about characteristics and roles of women and men pose a substantial problem by themselves – irrespective of the actual status quo – and thus require close examination.

Second, users of online discussion fora represent a certain selection of users, whereby the external validity of our findings is limited to that circle. However, given the global reach and growing importance of UGC as well as the public attention that vociferous actors from the online world receive, the population of users that we study is highly influential and thus of inherent relevance.

Finally, as argued above, we do not consider hate speech or open sexual harassment in our analysis but focus on more subtle forms of gender stereotypes. Although this limits the scope of our findings, we perceive it as a feature of our study: while it is relatively easy to detect gender discrimination in terms of open assaults and offenses, assessing subtle and subconscious gender stereotypes is way more difficult.[37] We provide an important contribution to addressing this challenge by proposing a novel classification procedure that allows us to document the prevalence and development of (subtle) gender

[37]In addition, focusing on subtle and subconscious gender stereotypes eliminates potential confounds regarding the supervision of online discussion forums. In particular, hate speech and open sexual harassment are often deleted by moderators. Since we discard such comments from our analysis, our results are unaffected by any potential moderation policies of the forum.

stereotypes over time.

Chapter 3

Machine Learning based Linkage of Company Data for Economic Research - Application to the EBDC Business Panels

3.1 Introduction

In the age of *Big Data*, the speed of information generation increases more and more. Thus, the possibilities for academics to use these data for economic research increase as well. Additionally, data get particularly valuable when different datasets with different kinds of info can be combined: Administrative data, survey data, proprietary data, and more can be linked to each other on a micro level to help to answer questions more timely, to correct for data errors, or to enable the study of completely novel questions. Thus, linking data is often an important task for economic research and policy advice. For example, Meyer and Mittag (2019) link survey and administrative data to overcome measurement error in household income, allowing for an improved evaluation of anti-poverty programs.

The process of linking entities from different data sources is called *Record Linkage* (RL) or sometimes entity resolution and it is straightforward when the data sources have a common unique identifier. Unfortunately, for German company data, there is not yet an agreed upon common identifier to enable such a linkage and this is said to be in part due to legal frictions (Neuscheler, 2023). Without a common identifier, records can still be linked via probabilistic matching: The more similar records are in attributes such as name or address, the higher the probability that they refer to the same entity (Fellegi and Sunter, 1969; Newcombe, 1988). Linkage errors, especially wrong matches, can introduce systematic measurement error, and thus bias, to downstream regressions that use linked data (Bailey et al., 2020). While the linkage of natural persons is already a nontrivial task[1], with records of non-natural persons there are additional complications: First, there is the hierarchical nature of companies where firms often belong to a group of firms, potentially with near identical name and addresses. Second, there are many changes that can occur over time such as reorganizations, name changes, or mergers.

The goal of this paper is to evaluate the use of Machine Learning (ML) and Natural Language Processing (NLP) methods for the linkage of company data. This is done

[1]Problems that can arise here are for example typographical errors, different spellings, nicknames, or name changes such as after a marriage (see e.g., Christen, 2012, p. 42ff).

at the example of the *ifo EBDC Business Panels* from the *LMU-ifo Economics and Business Data Center* (EBDC).[2] Here, I link the responding German firms from the long running surveys of the ifo Institue to their financial information from the commercial *Orbis* database via probabilistic matching. Orbis is an industry standard which is also used and linked for example by the research data centers of the *German Bundesbank* and the *IAB*, the research institute of the German Federal Employment Agency. While a linkage not based on ML has previously been conducted several years prior (Hönig, 2010), I now apply these newer techniques to achieve an improved match rate and re-evaluate older matches. This is motivated by the availability of more balance sheet records and because these methods can help overcome some of the challenges of company linkage.

For the linkage, I compute a matrix of various similarity metrics for pairs of records which I then use as an input for a supervised ML classification. To address the previously mentioned challenges of company data, I use comparison metrics that work well here and apply NLP methods which are uniquely applicable when dealing with company records: Because the words or *tokens*[3] in company names have a linguistic meaning, pre-trained embedding vectors (Mikolov et al., 2018) allow to extract this information.

The linkage results in a relatively high rate of matched entities, in particular for companies added more recently to the surveys. A substantially lower match rate for earlier decades could be due to past market exits or because there was more time for location changes or complex reorganizations. There also appears to be heterogeneity across sectors and surveys, with the construction survey having the lowest match rate. At the same time, the investment survey for manufacturing has a surprisingly high match rate despite the long survey run time, potentially due to firm characteristics like size. Matches with lower predicted match probability were manually corrected, revealing that false positives were almost exclusively cases where a firm was matched with a related entity like its holding. Linkage was particularly difficult when reorganizations within a corporate group occurred. This highlights that corporate structures and relations are a key challenge for company linkage.

This paper contributes to the literature introduced in section 3.2 by highlighting and addressing key challenges of company data linkage and giving some best practice advice. A second contribution is the evaluation of ML and in particular NLP methods for RL applications for steps beyond classification. I thus expand the growing literature of applications of ML methods for classification in applied linkages. A further contribution is that this paper serves as a documentation for the construction of the company correspondence table used for the final research datasets available at the EBDC.

The next section shows related literature and linkage applications. Then, section 3.3 describes the specific challenges one faces when linking company data and section 3.4 explains to what extent NLP methods can support here. Section 3.5 describes the data used for the linkage which is detailed in section 3.6. The results of the linkage are

[2]The EBDC then offers the resulting linked dataset for research at their premises. The EBDC is a Munich based accredited research data center at the ifo Institute and it provides secure access to company micro data for academic research at their workstations. Their well documented data include subjective micro data from the *ifo Business Surveys* alongside a version of this data enriched with companies' objective balance sheet data from *Hoppenstedt* and the *Bureau van Dijk* Databases such as *Orbis*. These linked datasets are called the *EBDC Business Panels*. Details about data access can be found in appendix section C.3. For more info see their website: `https://www.ifo.de/ebdc`

[3]The tokens of "Petra Mayer Sales GmbH" are "Petra", "Mayer", "Sales", and "GmbH".

then presented in section 3.7 and the discussion in section 3.8 lists avenues for further improvements. Finally, the paper concludes with section 3.9.

3.2 Related literature

The term *Record Linkage* is said to be coined by Dunn (1946), and Newcombe et al. (1959) proposed an automatic algorithm for linkage without common identifier based on agreement of other fields. These ideas were formalized by Fellegi and Sunter (1969) in an unsupervised framework that computes field specific match weights given how frequently pairs of records agree in the respective field. To determine matches, it then relies on an arbitrarily chosen cutoff for a similarity function that incorporates these weights. An advantage of this method is that it requires no training data. However, because the Fellegi-Sunter framework relies on rarely satisfied assumptions such as conditional independence of fields, supervised ML methods like support vectors machines, random forests and neural networks were instead proposed in other methodological papers (e.g., Tejada et al., 2001; Cohen and Richman, 2002; Bilenko and Mooney, 2003; Wilson, 2011; Schild et al., 2017; Cuffe and Goldschlag, 2018; Abowd et al., 2019)[4]

In recent years, the field of methodological RL research evolved further and Hettiarachchi et al. (2014) proposed a *next generation* of linkage using Neural Networks, genetic algorithms, and clustering methods. Thus, modern Deep Learning (DL) neural network architectures, like sequence models and convolutional neural networks are increasingly used for entity linkage (e.g., Gottapu et al., 2016, Ebraheem et al., 2017, Mudgal et al., 2018). Thanks to ML advancements, these applications can make use of *transfer learning*, where models are pre-trained on large datasets and can then be reused for various tasks with less training data. In particular, they make use of NLP methods in the form of pretrained language models, albeit not for company linkage but for example for products and bibliometric data. Mudgal et al. (2018) find that DL benefits only applications with textual or "dirty" data but not those with structured fields. However, by their definition, company names could be considered a dirty field where one can benefit from parsing its informational content using DL.

The particular challenge of linking business data and the need for further research in this field has already been acknowledged by Winkler (1995). However, the methodological literature on company RL appears to be smaller and focused on describing specific linkage cases[5] such as in Peruzzi et al. (2014), Schäffler (2014), Cuffe and Goldschlag (2018), Mason (2018), Moore et al. (2018), Schild (2016), Schild et al. (2017), Abowd et al. (2019), Gschwind et al. (2019), Eberle and Weinhardt (2020), and Doll et al. (2021). Likewise, the present paper is also focused on linkage methodology and serves as a major update to the linkage described in Gramlich (2008) by using more modern methods and technologies.

The original linkage did not use supervised ML but was instead closer to a variant of the Fellegi-Sunter framework. It relied on a set of very likely matches, the *gold standard*, identified via a simple heuristic, to compute field specific weights. This gold standard consisted of pairs that had identical phone numbers, fax number, or email addresses.

[4] A survey of the evolution of RL can be found in Binette and Steorts (2022)

[5] Potentially this is because there is a lack of standardized benchmark data.

However, since this information is often not available, a probabilistic linkage is still necessary. Thus, for other pairs, string similarity metrics for different fields were computed and aggregated in a linear combination with the field specific weights. A match decision was then made based on an arbitrarily chosen threshold on this linear combination. Thus, there are a few notable limitations of the original linkage: First, the previous linkage had a limited strategy to pre-select potential matches because it required an overlap in location information. This can introduce false negatives if the location is erroneously recorded. Instead, I opt for a combination of different pre-selection strategies that together can overcome some of their individual shortcomings. Second, the original linkage did not use ML with hand labelled training data for classification but relied on the set of ground truth pairs instead. A shortcoming of this approach is that whether or not phone and email address are present and overlapping may be nonrandom. Therefore, there can be selection into this gold standard set such that the computed weights may be less representative for other firms. Using hand labelled training data drawn at random, such as I do, alleviates this problem. Third, the previous linkage relied only on a single string similarity metric, whereas my approach employs different methods such that specific errors from individual metrics have a lower impact.

Applied empirical research shows the value of linked company data in economics: Gumpert et al. (2022) use data from administrative German social security records where employees' respective establishment is linked to firms from the Orbis database. This linkage allows to identify establishments belonging to the same firm to analyze how the managerial organization across establishments is interdependent for multiestablishment firms. Additionally, they can estimate how organization is affected by distance to headquarters due to geographic frictions. Aside from this, several papers previously used the *EBDC Business Panels*, i.e., the datasets that are being overhauled in this paper: For example, Huber (2018) analyzes the effect of bank lending cuts on firms and the local economy exposed to such cuts. Therefore, the author uses ifo survey information on the willingness of banks to grant loans and further matches this to a dataset about relationship banks from the credit rating agency Creditreform.[6] Furthermore, Enders et al. (2022) use the EBDC Business Expectation Panel to estimate the effect of firm expectations on later realized production and prices via survey questions. Here, they need the linked balance sheet data for propensity score matching to compare firms that have different expectations but the same fundamentals.

3.3 Challenges of company linkage

There are some general concerns that apply to any probabilistic linkage application such as tradeoffs between computational feasibility, accuracy, and coverage.[7] The key challenge is quality related since data can be outdated or wrong in both datasets. If records

[6]Firms are linked via the *Crefonummer*, a firm identifier that can be recovered from the balance sheet data source of the EBDC Business Panels.

[7]It is usually not computationally feasible to compare all entities of one dataset with all entities from another. Thus, to reduce the computational burden, practitioners need to make some assumptions about potential matches, thereby risking to make false negatives. The more restrictive this pre-selection, the more false negatives there can be, leading to a worse coverage. Likewise, increasing the accuracy may require more thorough comparison, thereby increasing computation time.

were perfectly maintained and clean, there would be no need for probabilistic linkage. Additionally, for RL supported by supervised ML, it is usually required to manually label training data but this is very time intensive and difficult. This difficulty comes from the fact that there are many factors such as historical changes to consider, often requiring close inspection of a company.

Linking non-natural persons such as companies comes with specific complications, in particular through (i) hierarchies, (ii) a lack of standards, and (iii) history which will be explained in the following:

Hierarchies Companies are hierarchical objects in two ways: First, firms can be part of larger corporate groups with separate entities for different functions. Consider the example of figure 3.1 where multiple companies belong to a *Petra Mayer* group and each entity appears in database *A* on the left.

Figure 3.1: Example for entities within a corporate group

Database A

Name	Address
Petra Mayer Sales GmbH	Abc-Str. 1
Petra Mayer Manufacturing GmbH	Abc-Str. 1
Petra Mayer Management GmbH	Abc-Str. 1

Database B

Name	Address
Petra Mayer GmbH	Abc-Str. 1

Note: Fictitious example of entities within in two databases.

These hierarchies can have horizontal elements, with different entities for example for producing and sales entities, and vertical ones, with management or holding companies. These entities can have very similar or even identical names and addresses. Additionally, there can be various reorganizations both within and across corporate groups due to acquisitions, mergers, fusions, internal activity shifts, renaming, or relocation. In fact, I find that a key source of false positives in company linkage can be a failure to identify the proper entity within a group rather than a link of completely unrelated entities. Figure 3.1 exemplifies this issue since it is not clear what the *Petra Mayer GmbH* of database *B* needs to be matched with from database *A* or whether this may be matched at all.

The second hierarchy related aspect is that the entities in different databases can be at different levels of aggregation. For example, Schild (2016) links establishment level to firm level data, Eberle and Weinhardt (2020) link two different data sources on establishment level, and Abowd et al. (2019) link employer information in a household survey to establishments. Entities of other levels of aggregation can appear with different names and addresses than their parent.

Antoni et al. (2018) further highlight that due to hierarchical complexities, constructing a final research dataset is nontrivial even when already provided with a correspondence table from an RL procedure.

Standards The company name is a collection of *tokens*, e.g., {*Petra, Mayer, Sales, GmbH*}, and it should ideally be a unique and common identifier. In reality, however, individual tokens can be excluded, included, or replaced across databases. This is a

concern because individual words might be crucial to differentiate entities within a corporate group. Additionally, even though company names shall have discriminatory power, Schild (2016) finds duplicate names for around 10% of companies in the Orbis database. Schäffler (2014) further identified that firms with identical names often belong to the same corporate group.

History If the identifying variables have been collected at different points in time across databases, the information can differ even if it contains no errors for example due to relocations or name changes. Furthermore, when dealing with panel data, it is possible that different entities would be a preferred match for different periods. This can occur, when a producing entity must be matched for its historical data but the corporate group has since moved its production to a different entity. Depending on the use case, different entities would then have to be matched in different periods.

Quality concerns are not unique to company linkage but they exacerbate above challenges. For example, sector information can be valuable to differentiate entities with different functions within a corporate group but it is not always clear how well it is maintained and how to best use it if many different sectors are specified.[8]

3.4 Natural Language Processing for company record linkage

Natural Language Processing (NLP) refers to techniques for computers to analyze natural language such as texts written by humans, for example to identify common patterns found in language. Some of these techniques use ML algorithms, for example for dimensionality reduction.

NLP techniques can support company linkage in ways that are not possible for linkage of natural persons because company names are made up of actual language words. For humans it is easy to understand, contextualize, and relate these but for machines to make use of this information, it needs to be processed by means of modern techniques.

One such method are word embedding vectors. They result from a dimensionality reduction technique where words are represented as fixed vectors in an n-dimensional space capturing semantic relations in language (Mikolov et al., 2013). These vectors are learned from a corpus of training data, for example by trying to predict the words surrounding a given word. Words used in similar contexts have similar embedding vectors because they need to be able to predict the same or similar surrounding words.[9] A variant of these particularly suited for RL is *FastText* (Bojanowski et al., 2017) because it is trained on pairs of characters rather than full words. This makes it more robust towards prefixes, suffixes, and even typos, all of which matter in RL. Because company names do not follow proper language rules[10], there are two limitations to using some modern NLP methods such as transformers where word vectors are context sensitive even in the

[8]Sector identifiers are also used for RL applications for example in Peruzzi et al. (2014).

[9]See for example Ash and Hansen (2022) for a description of these properties and applications in economic research.

[10]They are just a concatenation of words without proper context.

prediction stage: (i) It is not possible to infer word meaning from context and (ii) it is not possible to use the company names as training data because there is no proper context to learn from.[11] Both of these challenges can be overcome with *transfer learning*, i.e., by using embedding vectors pre-trained on other data. Other entity linkage research already used FastText transfer learning in Ebraheem et al. (2017), Mudgal et al. (2018), and Kasai et al. (2020), albeit not for company data but e.g., products and bibliometric data. I am using FastText embedding vectors pre-trained specifically for German language on massive text corpora of Wikipedia and Common Crawl by Mikolov et al. (2018)

Word embedding vectors can be used for various subtasks in a record linkage process: First, one can use them to compute similarity measures based on contextual similarity. To measure the similarity, I compute the *cosine similarity*[12] of two embedding vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ according to equation 3.1, where $\|\boldsymbol{v}\|$ and $\|\boldsymbol{w}\|$ refer to the euclidean norms of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ respectively.

$$similarity = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} \tag{3.1}$$

Table C3 in the appendix shows the three words most similar to a given word one might find in a company name. The method captures the meaning of words for example with sectoral, regional, or personal information well and can even enable to extract and standardize legal forms.

Second, word embedding vectors can be used to infer the meaning of words. Here, they allow segmentation of company names improving the quality of a linkage (see e.g., Christen, 2012 p. 55).[13]: (i) It reduces the linkage complexity when only tokens that serve the same purpose need to be compared and (ii) it allows for varying importance of types of words in a supervised classification step. For example, a token describing the industry helps differentiate companies with different roles in a corporate group. At the same time, differences in a legal form or appearing first name may be less relevant.[14] Figure 3.2 shows an example for segmented company names where the names are first split into individual tokens and then these tokens are assigned labels such as *legal form*.

Segmentation is achieved via supervised machine learning where a Neural Network sequence model with bidirectional LSTM nodes (Hochreiter and Schmidhuber, 1997) uses the FastText word embedding representation[15] of word as inputs to predict the label for each word.[16] The initial training data is taken from Loster et al. (2018)[17] and I iteratively

[11]For example, consider the terms *Sales* and *Management* in "Petra Mayer Sales GmbH" and "Petra Mayer Management GmbH". The training algorithm cannot infer that there is a different meaning between these two terms using just names as data. In actual language texts, however, these terms would likely appear with different surrounding words.

[12]This measure depends on the angle between the two vectors in the embedding vector space and is higher for words that are used in similar contexts in the training data. The cosine similarity of embedding vectors is also used as a similarity metric for entity linkage for example in Ebraheem et al. (2017).

[13]Also Loster et al. (2017) and Gschwind et al. (2019) show the usefulness of extracting and using company name segments, in particular *colloquial* names. Here, I extract the *proper name* (e.g., "Siemens") which should be identical to the colloquial name in many cases.

[14]For example because a name change can occur after heirs took over their parents' establishment.

[15]Transformation was supported with the *gensim* software library (Řehůřek and Sojka, 2010)

[16]This was done with *Tensorflow* (Abadi et al., 2015).

[17]They train a linear chain Conditional Random Field algorithm on this data for segmentation and do not make use of embedding vectors.

Figure 3.2: Example for company name segmentation

Original

Name
Petra Mayer Sales GmbH
ABC Gesellschaft mbH Germany

↓

Segmented

Proper Name	Person first name	Person Last name	Sector	Location	Legal
-	Petra	Mayer	Sales	-	GmbH
ABC	-	-	-	Germany	Gesellschaft, mbh

Note: Segmentation example for two fictitious company names.

expand this data by manually verifying or correcting predictions for German company names randomly selected from the Orbis database. The confusion matrix in figure 3.3 shows that the classification works well for most types [18] and most importantly for sector information, names of persons, locations and legal forms[19].

Another useful method proposed for RL in Cohen (2000) is the *term frequency - inverse document frequency* (TF-IDF), an information retrieval technique frequently used in NLP applications. It also transforms texts such as company names into real valued vectors in a high dimensional vector space. Here, a lower weight is given to tokens the more frequently they appear in other names[20] because the similarity of rare tokens is more informative than the similarity of a token shared by many company names (Spärck Jones, 1972).

3.5 Data

ifo Data The focal data consist of the contact information of participating firms[21] from the surveys of the ifo Institute and the main goal is to expand it with additional information about the respondents. Five surveys are considered: the monthly Business Surveys (IBS) for manufacturing (IBS-IND, 2019), retail and wholesale (IBS-TRA, 2019), construction (IBS-CON, 2019), services (IBS-SERV, 2019), as well as the biannual Investment Survey (IVS) for manufacturing firms (IVS-IND, 2019). All surveys regularly inquire business related information from German firms. Data from these surveys are used for example in Bachmann et al. (2013) who study the effect of uncertainty on firms' economic activity and Link et al. (2023) who compare firms' expectations and information

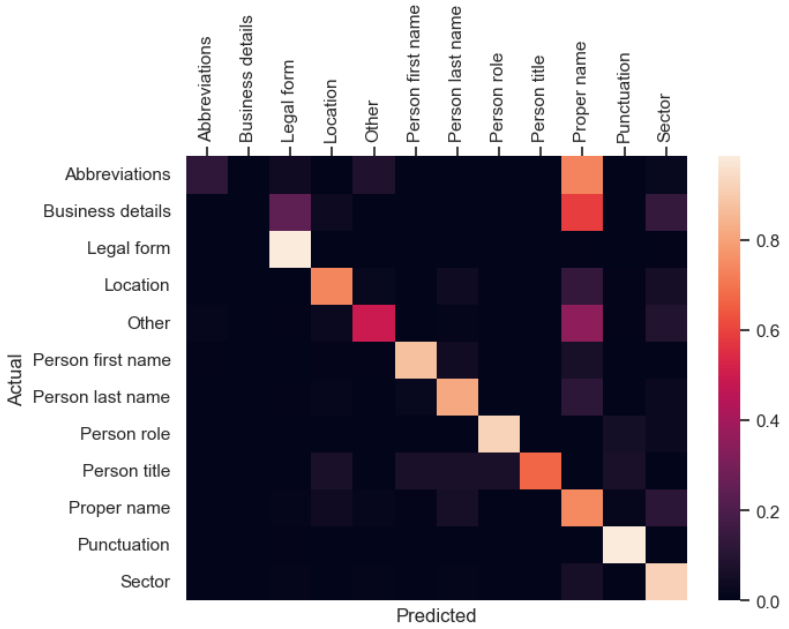
[18]The main errors are classifying some rarer words of the *business details*, *abbreviations*, and *other* as *proper name*.

[19]While legal forms are ultimately extracted from names via regular expressions, this shows that the method can be applied for this task as well.

[20]Additionally, the weight increases the more frequent a word appears within a company name. However, this property is less relevant in RL since words are rarely repeated in a name.

[21]This information is held separately from the survey responses and is otherwise not accessible for researchers using the survey micro data at the EBDC.

Figure 3.3: Confusion matrix for name segmentation



Note: The confusion matrix is based on cross validation using a held out portion of the labelled data. For each true label, it shows how many instances were predicted to be of each of the labels. The cell values represent shares of the row, i.e., of the actual labels and ideally these values would be all 1.0 on the diagonal indicating that there were no misclassifications. Brighter cells represent higher shares. *Business details* includes tokens such as *i.L.* (in liquidation). *Abbreviations* includes elements such as for example *BMW* as abbreviation for *Bayerische Motoren Werke* and can thus be easily confused with proper or colloqiual names. *Proper name* captures tokens such as *Siemens*, *Microsoft*, etc. *Location* includes words such as *Berlin*, *German*, and *International*.

frictions to those from the respondents of a household survey.

The ifo Institute conducts regular surveys since 1949[22] but earliest data are no longer available. Instead, data are accessible for the manufacturing sector from the 1960s onwards in the IVS and the 1980s in the IBS. Other sectors have been gradually added to the IBS, with the service sector being the most recent addition in 2001. The replies from the four sectors of the IBS enter the computation of the *ifo Business Climate Index*, a business cycle indicator for Germany. A complication with the IBS is that its different surveys have different levels of observations, either firms or products, and in some cases, this is not consistent over time.[23] Thus, a firm can submit multiple questionnaires within a given month. Questions of the IBS are more of a qualitative or subjective nature, where for example the business expectation is inquired with a three item likert scale. Table 3.1 contains the number of entities by survey and shows that manufacturing firms represent almost half of the more than 40.000 entities.[24]

There are several things to note: First, firms are not unique in the database because the IBS survey is partially on a product group level, because a firm can be engaged in

[22]The original goal was to provide information more timely than official statistics with the IBS being a monthly and the IVS a biannual survey (Sauer and Wohlrabe, 2020).

[23]see Link (2018) for more details.

[24]More detailed information about the surveys can be found in Sauer and Wohlrabe (2020).

Table 3.1: Number of entities in ifo survey database

Survey	Sector	Absolute	Share	Since	Frequency
IBS	IBS-CON	4797.0	0.11	1991	Monthly
	IBS-IND	12633.0	0.30	1980	Monthly
	IBS-SERV	7842.0	0.19	2004	Monthly
	IBS-TRA	9041.0	0.22	1990	Monthly
IVS	IVS-IND	7419.0	0.18	1964	Biannually
Total		41732.0	1.00		

Note: *IBS* is the ifo Business Survey and *IVS* is the ifo Investment Survey. *CON*, *IND*, *SERV*, and *TRA* respectively denote the surveys for the construction, manufacturing, services and retail/wholsale sectors. *Frequency* refers to the survey frequency, i.e., how often info is inquired from the participants.

multiple sectors, and because a firm can participate in both the IBS and the IVS.[25] It is furthermore possible that individual establishments of a company participate (Sauer and Wohlrabe, 2020). Thus, an *m-to-1* matching can occur, where different entities of the ifo database need to be matched with the same entity in the secondary database.

Second, especially before online submission became the dominant form of participation[26], the questionnaire was not necessarily always sent to the entity of interest. Instead, it is possible that for example the holding company collected questionnaires and forwarded them internally. In that case, the database contains the address and potentially name of the recipient rather than the inquired entity.

And third, the database may not always have been updated such that some changes may remain unnoticed. This is especially the case if a firm already ended participation or participates online, where no address is needed. Additionally, because the panel data shall be linked over time, it is possible that the entity of interest in the secondary database changes over time. This adds to the conceptual challenges mentioned in section 3.3.

BvD Data The secondary data source, i.e., the data I add to the primary source, is the commercial *Orbis* database from the publisher *Bureau van Dijk* (BvD).[27] It contains objective quantitative financial and other information such as balance sheets, patents, or shareholder structures as opposed to qualitative survey responses. BvD receive their data for this global database from various sources such as for example *Creditreform Rating AG* for Germany. The entities are on a company level with some of them being so called *branches* of other companies.[28].

The previous linkage of ifo data described in Gramlich (2008) relied on entities from the *Amadeus* database, another BvD product with a focus on European firms. It is supposed to be a subset of Orbis with the same ID numbers. Because it is relatively

[25]Nonetheless, these duplicate firms each have unique ID numbers.

[26]By now, around 60% participate via the web (Sauer and Wohlrabe, 2020).

[27]The company databases from *Bureau van Dijk*, have already been used in RL applications for example in Peruzzi et al. (2014), Schild (2016), and Schild et al. (2017).

[28]These are easily identifiable because they have the same ID number as the unit they branch off of but add a running number as suffix such as for example DExxxxxxxxxx-1000. Table C1 shows that around 3.6 percent of IDs in the data are such branch IDs.

common for institutions to link their data to Orbis, the BvD ID can then also allow for subsequent linkages with other micro data e.g., from other surveys.

I use the *2018-06* snapshot from the Orbis flatfiles and export all German companies. I further remove all companies with an ID not starting with *DE* (0.2%) and those whose ID starts with *DE** (1.3%)[29]. This leads to a final selection of 3,759,447 unique BvD IDs.[30]

Available information I use company names, address, sector identifiers, and other contact information such as telephone numbers. These variables are not always directly comparable and need some standardization. For the name, the ifo data store only a single variant, whereas BvD data also contain previous company names and *also-known-as names* if applicable. Likewise, for the sector information, the primary data source (ifo) contain only a single 4-digit sector identifier[31] and the BvD data contain multiple secondary NACE Rev. 2 sector identifiers.

Table 3.2 shows the share of missing information from both data sources after the preprocessing steps described in section 3.6.1. By design, the company name is always available, since only entities that provide one are considered. Address information is also available in at least 93% of cases. Additional contact information such as phone and fax numbers are frequently missing and the email address is missing in the majority of cases. Thus, these variables are difficult to use for some applications like indexing.

Table 3.2: Share of missing data

	Orbis	ifo
Name	0.00	0.00
Federal state	0.06	0.00
City	0.04	0.05
Address	0.07	0.05
Street	0.07	0.05
Postcode	0.05	0.05
Address number	0.08	0.08
Sector 1 digit	0.16	0.01
Sector	0.16	0.30
Phone	0.55	0.12
Fax	0.68	0.21
Email	0.70	0.52

Note: Share of missing observations after the preprocessing procedure which includes removal of fields that appear erroneous and filling of some missing fields. Thus, for the ifo data, the one digit sector is more frequently available than the full number because it can frequently be inferred when the firm participates in the trade or construction surveys.

[29]These do not contain any relevant financial info and are an artifact of entities that could not be matched to other existing IDs by BvD.

[30]This includes both companies registered in the commercial register and unregistered traders. It also includes the *branches*. Additionally, these companies can be active or inactive. Because historical data shall be linked, inactive companies are relevant as well.

[31]Depending on the survry, this is either the WZ08, the WZ03 or the WZ93. During the preprocessing described in section 3.6.1 these are unified as well as possible to the WZ08.

3.6 Record Linkage procedure

The linkage follows a typical workflow as described for example in Christen (2012), a general guidebook for practitioners of RL. Five steps are required here: (i) *Preprocessing* to make the records as comparable as possible. (ii) *Indexing* to narrow down the search space and determine a set of pairs to consider. (iii) *Comparison* to compute a vector of similarity metrics for each pair. (iv) *Classification* of pairs as matches or non-matches using their similarity vector as inputs. And (v) *postprocessing* which includes filtering out ambiguous matches. The rest of this section describes these steps in detail.

3.6.1 Preprocessing

Even though the linkage is designed to overcome errors and differences in the datasets, it is important to facilitate this by preparing and cleaning the records of both data sources. The main tasks here are (i) standardization, (ii) filling missing information, (iii) feature generation, and (iv) transformation.[32]

Standardization serves to make the records comparable across databases. This includes case folding and replacing German *Umlaute* ä, ö, ü with a, o, and u respectively. Additionally, legal forms are extracted from company names via regular expressions which are slightly adjusted from Schild et al. (2017). In Orbis, there can be multiple variants for the name, city, address, phone number, and fax number, any of which could be found in the other database. For this reason, I store these alternatives into sets[33] that allow for comparison via set methods as described in section 3.6.3.

Filling missing data is particularly relevant for the indexing step where an exact overlap in some field of choice is required. If info in the respective indexing field is missing, the record can not be linked to any other record. Filling information with the help of other fields is possible only in few specific situations. A table containing all German zip codes, their respective municipality, and other regional information (Deutsche Post Direkt, 2019) allows to infer the zip code from the location or vice versa if uniquely possible. The different ifo surveys have different sector identifiers[34] such that they need to be harmonized to the WZ08 industry classification which is roughly equivalent to the NACE Rev. 2 available in Orbis. This is achieved using WZ03 to WZ08 correspondence tables (Destatis, 2008) and in some cases, a one-digit identifier can be inferred from the survey sector itself.[35]

Feature generation infers new *features* – the equivalent of *variables* in inferential statistics and econometrics – i.e., attributes for a machine learning model, from available data: The zip code identifies the federal state and the four digit sector identifier can be aggregated into more coarse categories.[36]

Transformation creates new fields as transformations of existing ones. Here, I use

[32]The steps were in part inspired by Schild (2016). The full list of measures can be found in appendix section C.2.1.

[33]E.g., phone numbers: {+123 456789, +987 654321}.

[34]See Link (2018) for a very detailed overview.

[35]E.g., all entities in the construction survey should have a "4" as first digit.

[36]It is possible to aggregate to 3-digit, 2-digit, 1-digit, to only a differentiation between *manufacturing*, *trade*, *construction* and *service* sector. This is helpful because the more coarse the information, the more likely it is that a true match agrees.

Phonetic encoding to counter different spellings. I encoded attributes with the *Double Metaphone*[37] encoding (Philips, 2000) which is designed to work with a number of different languages, including German. This transforms for example the word "Maschinenbau" (engineering) to "MXNNP"[38]. Because, by removing important differences, phonetic encoding can worsen match rates when used for comparison (Bailey et al., 2020), I only use it for selecting candidate pairs in the indexing step. Additionally, I use the FastText NLP method introduced in section 3.4 by transforming the *city* field into its embedding vector representation and by segmenting the company names. Due to data protection concerns, ifo data need to be processed on a specially protected computer, whereas the Orbis data could be preprocessed on a different machine. Because of hardware limitations of the protected device, the segmentation could only be executed for the Orbis data on a different computer. Thus, rather than comparing the same tags[39] of both datasets, I check whether there is an overlap between each tag of the Orbis segments with all of the tokens of the ifo company. Under the assumption that a segmentation of the ifo data would have resulted in the same labels for the same words, this second best approach should not differ much from the optimum. This is plausible because the segmentation relies mostly on the fixed word embedding vectors of company name tokens such that the same words are likely predicted equally in both data sources.

The result of the preprocessing step are two tables, one for each data source, with the cleaned contact information of firms.

3.6.2 Indexing

The set of all possible pairs is the cartesian product of both data sources, i.e., of size $n \times m$ with n and m respectively being the number of records in both sources. Even if one of the datasets contains just tens of thousands of observations, the computational cost of this set can be prohibitively large when the other dataset has millions of records as it is the case for the Orbis dataset. Thus, the indexing step serves to select a set of potential pairs to consider for further linkage steps. Figure 3.4 shows an example where the number of pairs is reduced from $3 \times 3 = 9$ to 3.

The vast majority of potential pairs is no match and is very dissimilar such that it makes sense to filter them out using a fast selection method in the form of *blocking* (Newcombe et al., 1959; Newcombe and Kennedy, 1962) and *filtering* first.[40] Blocking requires pairs to perfectly agree on a set of predetermined fields, the *blocking keys*.[41] For example, records can be required to have the same sector code. Filtering is a step applied after the blocking and it requires records to have some minimum similarity score in a

[37] I used a python implementation from
https://github.com/dracos/double-metaphone/blob/master/metaphone.py

[38] Some words can be encoded into a primary and a secondary encoding. In this case, only the primary encoding is utilized.

[39] The *tag* refers to the label of tokens. I.e., when comparing sector tags, this refers to an list of all words with sector information within a company name.

[40] The usage of both methods together is for example suggested by Papadakis et al. (2019).

[41] Additionally, I use *Sorted Neighborhood Blocking* (Hernández and Stolfo, 1995) which makes the indexing more robust to noisy data (Papadakis et al., 2019). Here, records are sorted on a predetermined key and rather than requiring a perfect overlap, a fixed size window is moved over the records such that all records that lie within this window are considered as pairs.

Figure 3.4: Indexing example

Database A

ID_A	Name	Street
A1	Petra Mayer Sales GmbH	Abc-Str.
A2	ABC Gesellschaft mbH Germany	-
A3	XYZ AG	Ghi-Str.

Database B

ID_B	Name	Street
B1	ABC GmbH	Def-Str.
B2	Petra Mayer GmbH	Abc-Str.
B3	Maier GmbH	Xyz-Str.

↓

Indexed pairs

ID_A	ID_B	$Name_A$	$Name_B$	$Street_A$	$Street_B$
A1	B2	Petra Mayer Sales GmbH	Petra Mayer GmbH	Abc-Str.	Abc-Str.
A1	B3	Petra Mayer Sales GmbH	Maier GmbH	Abc-Str.	Xyz-Str.
A2	B1	ABC Gesellschaft mbH Germany	ABC GmbH	-	Def-Str.

Note: Fictitious example of the indexing step. The tables of database A and B each contain records' cleaned attributes after the preprocessing step. Only the indexed pairs are considered for further linkage steps. The table of indexed pairs contains the attributes of the records from both data sources. The index itself refers to the columns ID_A and ID_B from the table of indexed pairs.

given field. The similarity measure is ideally fast to compute such that it can be done for a larger set of pairs.

Both methods come with a trade-off: one the one hand, stricter rules make the search computationally feasible and on the other hand, stricter rules can lead to false negatives, for example when there are errors in the blocking key. To mitigate this concern, it is suggested to use the union of pairs from multiple different blocking and filtering strategies as final index (Herzog et al., 2007).[42] The basis of most strategies are combinations of sector or location based blocking keys as these are frequently filled. Additionally, the respective Orbis-*branches* of candidates and previously collected ML training data pairs were included in the index.

We have already seen in section 3.3 that the temporal dimension can introduce challenges for panel data sources. Potentially, an ifo ID needs to be matched to one Orbis ID for older historical information and to a different one for more recent observations for example due to a restructuring of the company. Here, I propose to do two separate linkages, a *pre* and a *post* linkage: The *pre* linkage considers only pairs where the date of incorporation from Orbis was before the ifo survey start, i.e., the company must have existed when it participated in the survey. Conversely, the *post* linkage considers only pairs where the date of incorporation was *after* the survey start, i.e., this entity was founded at some point in time during the survey runtime.[43] The following steps, i.e., comparison, classification, and postprocessing, are then all conducted separately for both the *pre* and *post* pairs.

Table 3.3 shows the blocking and filtering keys by strategy. The union of all strategies leads to a final index of around 4.4 million unique pairs. This is substantially larger than

[42]Both the indexing and comparison step were mostly executed with the *Record Linkage Toolkit* (De Bruin, 2019) with additional metrics from the *jellyfish* and *textdistance* packages in python.

[43]Thus, the comparison of the date of incorporation on the one side to the year of survey start on the other can be seen as a *complex feature* according to Wilson (2011).

the index any single one of these strategies would achieve but still just a small fraction of the more than 120 billion pairs of the full index.

Table 3.3: Indexing strategies

Strategy	Block	Filter	Pairs pre	Pairs post	Combined
1	plz (5d), legal, extra	sector (1d, multi): exact, name tokens (DM): jaro $\geqslant$ 0.9	106,122	14,108	120,230
2	plz (4d), sector (section), extra	name tokens (DM): jaro $\geqslant$ 0.8	756,962	125,804	882,766
3	plz (5d, sorted N=3), legal, sector (group), extra	add. number range: exact, name tokens (DM): jaro $\geqslant$ 0.8	40,415	3,833	44,248
4	city (DM), legal, sector (4d, sorted N=7)	street (DM): jaro $\geqslant$ 0.7, name tokens (DM): jaro $\geqslant$ 0.8	153,480	7,371	160,851
5	email, sector (section), extra	name tokens: exact	567,661	64,002	631,663
6	fed. state, legal, extra, city (DM, sorted N=3)	name tokens (DM)	1,309,197	185,471	1,494,668
7	city, address number	sector (2d, multi): exact, name tokens (DM): jaro $\geqslant$ 0.9	29,860	3,303	33,163
8	street (DM), plz (3d, sorted N=7)	name tokens (DM): exact	138,800	25,814	164,614
9	sector (section), legal, extra, plz (3d, sorted N=3)	city: jaro $\geqslant$ 0.9, name tokens (DM): exact	302,188	38,704	340,892
10	sector (section), legal, extra, plz (3d, sorted N=3)	street (DM): jaro $\geqslant$ 0.8, name tokens: jaro $\geqslant$ 0.8	58,392	9,382	67,774
11	fed. state, email (sorted N=7)	name tokens (DM): jaro $\geqslant$ 0.8	352,685	42,489	395,174
12	city (DM), extra, street (DM, sorted N=5)	name tokens: exact	100,930	18,446	119,376
13	street (DM), extra, city (DM, sorted N=3)	name tokens: exact	92,261	15,376	107,637
Extra pairs			782,851	100,685	883,536
Total			3,834,242	533,245	4,367,487

Note: Table shows the blocking and filtering keys for different indexing strategies. *Block* refers to the blocking keys, where an exact match of the entire variable is required for pairs to be considered. *Filter* refers to the variables for the filtering step conducted after the blocking, where a simple comparison metric is computed and pairs are required to have a minimum similarity in this metric or partial overlap in the variable. For computational reasons, an exact overlap in one token of an array variable, here indicated with *multi*, is computed in the filtering rather than the blocking step. Omitted from this table is the additional filtering that separates the *pre* from the *post* linkage which is based on a comparison of the *date of incorporation* from Orbis and the *survey start date* from ifo. *Extra pairs* contains pairs from existing training data pairs, some previous matches and the *branches* from Orbis IDs. *Sorted N* refers to sorted neighborhood matching and the number indicates the window size. *DM* refers to a variable phonetically encoded with with double metaphone. *Name tokens* contains the set of all name tokens from all name variants. *1d, 2d, ...* respectively refer to the number of first digits. *plz* refers to the postcode. *legal* refers to the legal form.

The *pre* and *post* indexing steps each result in a correspondence table with the ID numbers of considered pairs.

3.6.3 Comparison

The basis for classifying the candidate pairs from the indexing step as matches or non-matches is the matrix of their similarity scores. Cuffe and Goldschlag (2018) suggest that linkages can be more effective by combining many different comparison metrics. The full list of comparison metrics is shown in appendix table C4. Figure 3.5 exemplifies this step based on the example index from figure 3.4.

For this study, I choose methods that I expect to work well with the specific challenges of company data: (i) Order robust string comparison methods, (ii) array methods, (iii) TF-IDF based methods, and (iv) embedding methods. Order robust string comparison metrics are useful for company names because they compute the similarity between two strings such that the order of tokens has less impact. Here, I use Longest Common Subsequence (LCSSeq) (Hirschberg, 1977), Longest Common Substring (LCS) (see e.g., Gusfield 1997), Character n-gram similarity[44] (Ukkonen, 1992), Cosine similarity of character n-grams, and Smith-Waterman (Smith and Waterman, 1981). Array or set

[44]This method is commonly used for company names, for example in Gramlich (2008) and Schild (2016).

Figure 3.5: Comparison example

Indexed pairs

ID_A	ID_B	$Name_A$	$Name_B$	$Street_A$	$Street_B$
A1	B2	Petra Mayer Sales GmbH	Petra Mayer GmbH	Abc-Str.	Abc-Str.
A1	B3	Petra Mayer Sales GmbH	Maier GmbH	Abc-Str.	Xyz-Str.
A2	B1	ABC Gesellschaft mbH Germany	ABC GmbH	-	Def-Str.

↓

Similarity matrix

ID_A	ID_B	$ngram_{Name}$	LCS_{Name}	$Jaro_{Street}$
A1	B2	0.739	0.842	1.000
A1	B3	0.391	0.562	0.750
A2	B1	0.276	0.444	0.000

Note: Fictitious example of the comparison step. Here the attributes such as name or street of pairs which given by the index from the indexing step are compared with string similiarity metrics. Thus, $ngram_{Name}$ refers to the ngram similarity between $Name_A$ and $Name_B$. Similarities are computed on the raw strings for this example. Results in the actual linkage are likely more favourable thanks to the cleaning from the preprocessing step which is omitted here for simplicity.

methods are those that compare sets of tokens, e.g., the list of all words in a name, rather than strings and they come in different forms: First, it is possible to check if two records have any overlapping tokens or compute the share of overlapping tokens.[45] Thus, one can check for overlapping tokens between {*"Petra", "Mayer", "Sales", "GmbH"*} and {*"Petra", "Mayer", "GmbH"*}. Second, one can compute string similarities for all possible combinations of token pairs across two sets to get the maximum similarity or compute a fuzzy overlap where it is sufficient for tokens to have a minimum string similarity for a binary overlap indicator[46]. For sectors, I use array methods with information on all available sector identifiers provided by Orbis. Relying only on the main sector could result in an overrepresentation of companies which are mostly active in their main sectors and low match rates for survey responses about other activities (Schild, 2016. Array methods are also applied to the company name segments where, for a subset of segment categories[47], the overlap of Orbis tokens from this category with all ifo tokens is computed. To weight tokens based on their relative frequency, I apply both cosine similarity on TF-IDF vectorized record fields and Soft TF-IDF. The latter is a measure that often performs very well in RL applications (Cohen et al., 2003) by combining string similarity with frequency weights to also consider similar tokens. Embedding methods use the cosine similarity between embedding vectors of tokens as described in section 3.4. This helps identify tokens that are typically mentioned in similar contexts. Here, I use

[45]Further set methods I utilized were the cosine similarity between words, the Jaccard index (Jaccard, 1912), and Monge-Elkan (Monge and Elkan, 1996).

[46]The latter is used only for filtering in the indexing step. With this filter, pairs are required to have a token-wise Jaro similarity (Jaro, 1989) of 0.8 or 0.9, depending on the attribute.

[47]Here, I restrict the analysis to the segment categories *location*, *person first name*, *person last name*, *sector*, and *proper name* because these were well classified and I expect them to be the most useful in separating companies.

them for location information[48] to capture for example similarities in locations of different geographic hierarchy. Figure C2 shows that the cosine similarity of location info word embeddings is highly correlated to string based similarity metrics. At the same time, it is the least correlated among all the metrics, implying that it may capture some additional information.

The comparison step results in a matrix[49] where each row is the vector of similarity metric scores for a given considered pair of records.

3.6.4 Classification

The comparison metrics for the *pre* and *post* candidate pairs allow to differentiate between matches and non-matches. To this end, I use the comparison matrix as input to a supervised ML classification[50] with manually labelled record pairs as training data. Figure 3.6 shows this process for the exemplary similarity matrix from figure 3.5.

Figure 3.6: Classification example

Similarity matrix

ID_A	ID_B	$ngram_{Name}$	LCS_{Name}	$Jaro_{Street}$
A1	B2	0.739	0.842	1.000
A1	B3	0.391	0.562	0.750
A2	B1	0.276	0.444	0.000

→

Classification

ID_A	ID_B	Match probability
A1	B2	0.72
A1	B3	0.35
A2	B1	0.31

Note: Fictitious example of the classification step. The match probabilities in this example are not generated by the model and only serve for illustrative purposes.

As suggested for example in Bailey et al. (2020), the algorithm for classification is an ensemble of several different estimators, each with different transformations and comparison metrics. Table 3.4 lists the individual models that make up the ensemble. A stratified 10-fold cross validation helps tuning the hyperparameters of the individual models and their preprocessing pipelines. The final score is aggregated using a logistic regression that takes predicted probabilities of the ensemble models as input. Rather than using all available comparison metrics as inputs, most models use only a subset of features[51] or reduce dimensionality via *principal component analysis* (Pearson, 1901; Hotelling, 1933).

Because the vast majority of pairs are non-matches and because of the bimodal similarity distribution, selecting pairs to label at random can result in a set of many completely dissimilar pairs and a few very similar ones. This leads to a few challenges: First, a classifier might opt for always predicting non-matches if one does not otherwise address this imbalance. And second, the pairs in the training data likely consist only of extremes and there is no support for the more difficult cases such that the method cannot learn such patterns. Because labelling is very time consuming, it is not feasible to draw and label a random sample with sufficient support for all the different cases. Thus I opt for

[48] In a follow up study, this method could also be well applied to the company name, in particular to the name segment that contains sector information.

[49] One for both the *pre* and *post* linkages respectively.

[50] Training and prediction were done using the *scikit-learn* library (Pedregosa et al., 2011).

[51] Features are selected via an aggregation function, via penalized regression, or via some heuristic.

Table 3.4: Components of the supervised ML ensemble

	Estimator	Description
1	LogisticRegression	feature aggregation (max), continuous features
2	LinearSVC	feature aggregation (max), continuous features
3	MLPClassifier	feature aggregation (max), continuous features
4	XGBClassifier	feature aggregation (max), continuous features
5	LogisticRegression	feature aggregation (mean), continuous features
6	LinearSVC	feature aggregation (mean), continuous features
7	MLPClassifier	feature aggregation (mean), continuous features
8	XGBClassifier	feature aggregation (mean), continuous features
9	LogisticRegression	frequency weights, no missing data indicators
10	LinearSVC	frequency weights, no missing data indicators
11	MLPClassifier	frequency weights, no missing data indicators
12	XGBClassifier	frequency weights, no missing data indicators
13	LogisticRegression	continuous features
14	LinearSVC	continuous features
15	MLPClassifier	continuous features
16	XGBClassifier	continuous features
17	RandomForestClassifier	categorical features, binned continuous features
18	CatBoostClassifier	categorical features, binned continuous features
19	RandomForestClassifier	no missing data indicators, binned continuous features
20	CatBoostClassifier	no missing data indicators, binned continuous features
21	MLPClassifier	frequency weights, categorical features
22	LogisticRegression	PCA
23	LinearSVC	PCA
24	MLPClassifier	PCA

Note: Overview over the individual models that enter the ensemble. Each model has its own pipeline with various transformation and selection steps such as aggregation of all location-based features. These steps are shown in the description column. *MLPClassifier* is a Multilayer Perceptron, i.e., a Neural Network. *XGBClassifier* and *CatBoost* are Gradient Boosting classifiers with the latter supporting categorical features. *LinearSVC* refers to a Support Vector Machines algorithm with a linear kernel.

an iterative *active learning* approach, where I draw labelling data given their predicted match probabilities from a previous iteration which depended on fewer training instances. An active learning approach is also suggested and used for RL in Tejada et al. (2001), Sarawagi and Bhamidipaty (2002), Isele and Bizer (2013), Qian et al. (2017), and Kasai et al. (2020). With this approach, I can ensure that the number of matches and non-matches is more balanced and at the same time, I oversample difficult cases by drawing relatively more pairs with a predicted match probability of around 30-70%. Appendix table C5 shows how many instances are in the training data. *Training data 1* (8,307 instances) is used to train the individual models of the ensemble and *training data 2* (3,561 instances) is used to train the ensemble aggregator.

A drawback of this active learning approach with oversampling of difficult cases is that it is not straightforward to evaluate the algorithm with an unbiased performance metric. Nonetheless, for transparency, I include table C6 with the classification metrics for each model of the ensemble in the appendix. This table also highlights that the ensemble outperforms any of its individual components with both a comparably high recall and precision.

The classification results in a vector[52] containing the predicted match probabilities for each of the considered pairs.

3.6.5 Postprocessing

Because the classification can result in a many-to-many matching, a postprocessing step ensures there is only one Orbis ID per ifo ID.[53] Table 3.5 shows that for around one quarter of the ifo IDs with any match, there is more than one match. Thus, for each ifo ID, I keep only the match with the highest predicted match probability.

Table 3.5: Multiple matches per ifo ID in the pre linkage

Matches per ifo ID	1	2	3	4	5	6+
Absolute	22210	5243	1230	330	119	134
Share of IDs	0.76	0.18	0.04	0.01	0.00	0.00

Note: This table shows that, before postprocessing, matches are not unique per ifo ID for around 24% of entities. Hence, this needs to be reduced to one BvD match per ifo ID. These values are not yet indicative of the final match rate after postprocessing. See section 3.7 for that. *Share of IDs* refers to the share of matched ifo IDs and it sums up to one with some rounding imprecision.

Ultimately, a manual review of the remaining matches allows to correct for mistakes. To avoid systematic errors in downstream analyses, it is important to avoid false positives more so than false negatives (Bailey et al., 2020). For this reason and because the review is very time consuming, it is limited to correcting for false positives with a predicted match probability in the range between 50% and 90%[54]. Additionally, some pairs in the range from 40% to 50% were manually corrected as well to evaluate the extent of false negatives and increase the recall. Figure C3 shows the share of corrected entities by predicted probability in the *pre*-linkage. The error rate at 90% was very low and entities from the manufacturing surveys needed to be corrected the most. The errors made by the classifier are almost exclusively cases where the wrong company within a corporate group has been selected or where it was not possible to manually label a match with certainty[55].

The postprocessing step results in one correspondence table with the matched ifo and BvD ID numbers for both the *pre* and *post* linkages respectively.

3.7 Results

The linkage results in two correspondence tables: A larger *pre* table, containing only Orbis companies founded *before* the firm started to participate in the survey and a smaller *post*

[52]Actually two vectors, one for the *pre* and *post* linkages respectively.

[53]A single Orbis ID can belong to different ifo IDs if that company participated in multiple surveys at one point in time.

[54]The distribution of probabilities is bimodal with its peaks on the two extremes, i.e., very low and very high probabilities. The middle on the other hand contains comparably few observations such that manual review is feasible. Going beyond 90% is impractical due to the high volume and quality of match pairs.

[55]This occurs when it seems ambiguous which match candidate is the correct one.

table, containing only Orbis companies founded *after* survey participation.

Match rate Table 3.6 shows how many ifo IDs are be matched in each survey. The majority of companies has a match and these are primarily coming from the *pre* linkage, as expected. The number of matches from the *post* linkage is smallest for the service sector survey, with only 20 identified matches, and largest for the two manufacturing industries surveys. A possible explanation for this is that the IBS-SERV started in 2004, while data for the IVS-IND and IBS-IND are available since 1964 and 1980 respectively. In such a long time span, reorganizations are more likely. Also other factors influence the match rate: Despite the long time the survey has been running, IVS-IND has the second best match rate. This may be explained for example by the nature of the manufacturing companies in this survey. Here, larger companies are more strongly represented than smaller ones (Sauer and Wohlrabe, 2020) and thus there can be a lower risk of these entities exiting the market (Aldrich and Auster, 1986). Appendix table C2 confirms this by showing that the manufacturing companies in the surveys tend to be larger. On average, the IVS-IND firms have almost seven times as many employees as the IBS-CON firms, potentially explaining why the construction companies have the worst match rate with 65% of IDs matched. Sector differences can also be driven by organizational differences or naming conventions. For example, figure C1 in the appendix displays name changes recorded in the Orbis database by sector and shows that construction companies are subject to substantially more name changes than manufacturing or trade companies.

Table 3.6: Match rates by survey

Survey	All ifo ids	Matches pre	Matches post	Share of ifo ids with any match
IBS-CON	4797	3074	104	0.65
IBS-IND	12633	8486	395	0.69
IBS-SERV	7842	6187	20	0.79
IBS-TRA	9041	6154	147	0.69
IVS-IND	7419	4813	796	0.72

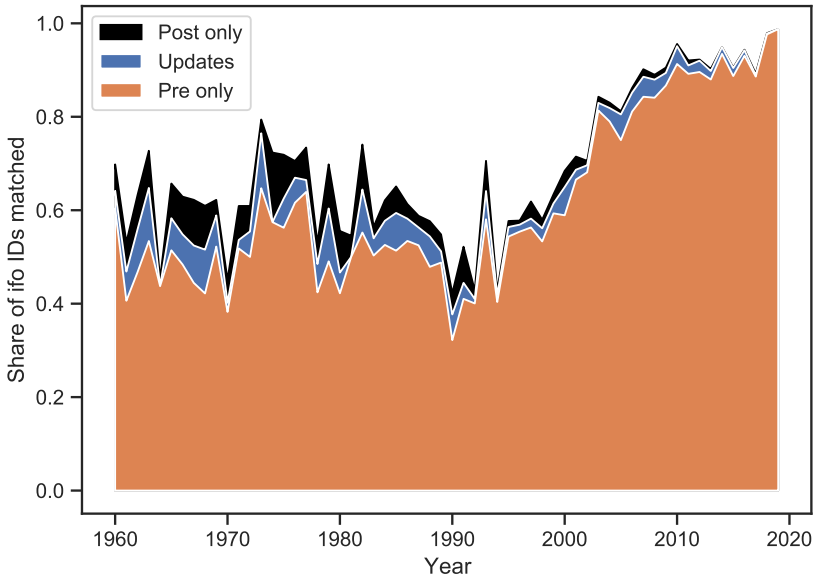
Note: Table shows the matchrate, i.e., the share of ifo IDs that could be matched to an Orbis entity. *Matches pre* is the number of ifo IDs matched in the *pre* linkage. *Matches post* is the number of ifo IDs matched in the *post* linkage. *Share of ifo ids with any match* is the share of unique ifo firms matched in the pre, post, or both linkages.

Figure 3.7 supports the hypothesis that it is primarily older entries that receive updates in the post linkage. It also shows that the linkage rate improves over time and nearly all firms added to the survey in the most recent years can be linked. One can also see that a substantial fraction of IDs from the early years has only a match in the post table. This occurs for example when the original firm ceases to exist after a reorganization and is not listed in Orbis.[56]

A multivariate regression allows to analyze this more systematically by showing how the match rate varies with different attributes holding all others fixed. Figure 3.8 shows the coefficients from a regression of the match status on various observed firm characteristics: Despite their high overall linkage rate, the linkage appears to be most difficult

[56]It is not clear when and under which conditions such inactive firms are not listed in Orbis.

Figure 3.7: Matchrate by year of survey start



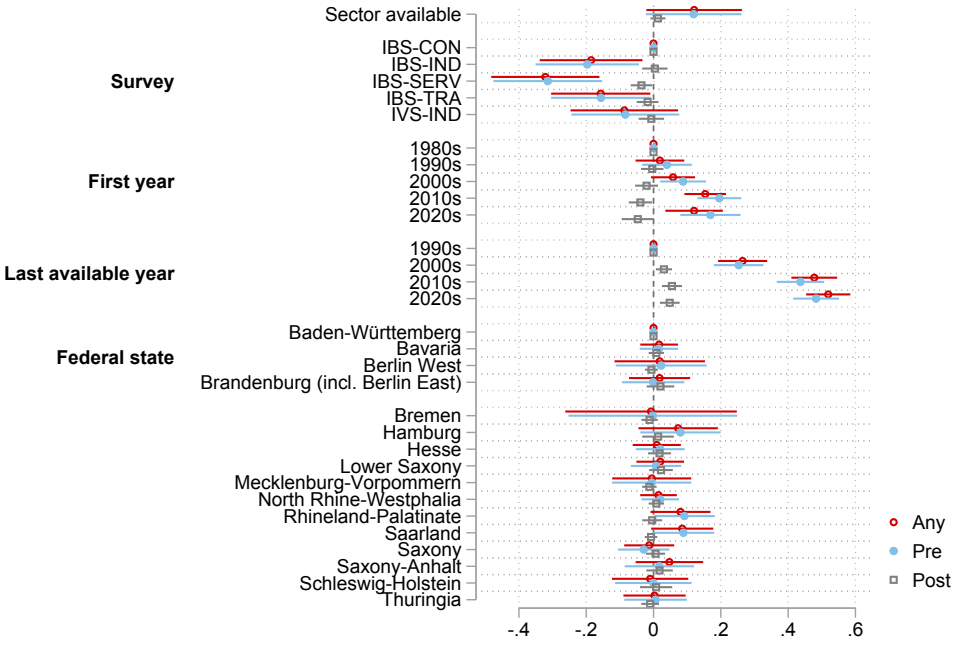
Note: The figure shows the matchrate, i.e., the share of ifo IDs that could be matched to an Orbis entity. The x-axis represents the year an entity was added to the survey and does not need to coincide with its year of incorporation. *Pre only* refers to ifo IDs which could only be linked in the *pre* linkage, i.e., to a company that existed before the survey start. *Post only* refers to ifo IDs which could only be linked in the *post* linkage, i.e., to a company founded after the survey start. *Updates* refers to ifo IDs which could be linked to different entities in both the pre and post linkages.

for service companies when conditioning on other factors. Furthermore, companies from eastern Germany have a lower matchrate than those from western Germany. A very strong predictor of matchrate is when companies still participate or ended participation just recently, likely because this indicates a higher probability that the firm address information is still up-to-date. There are also some substantial differences between the pre and post linkages: Post linkage rates are higher for eastern Germany and they decrease for more recently added entities. The latter is intuitive since these likely did not experience an organizational shift in the shorter period. Additionally, the coefficients on employment size range dummies are reported in appendix figure C4. They show that the linkage appears to be easier for medium sized companies, while very large companies are harder to match. A potential reason for this is that larger companies can be organized in more complex corporate groups.

Metric importance With the variety of different comparison metrics used, it can be helpful to see which of these are particularly useful in finding matches. This allows to narrow down the metrics and thus decrease computational cost in future applications. Because there are different algorithms in the ensemble and they use different sets of features, it is necessary to evaluate feature importances with a model agnostic framework.

One such method is the recent *SHAP* algorithm by Lundberg and Lee (2017) which they introduced for more interpretability of modern black box prediction models. It does so by computing values informative about the importance of each feature for a given prediction or set of predictions. The method is based on the Game Theoretic concept of

Figure 3.8: Regression of match status on firm characteristics



Note: Coefficients of a regression of match status on characteristics from the ifo companies. Employment size range dummies are also included in the regression but their coefficients are only shown in appendix figure C4 for better visibility. Point estimates of a linear probability model are shown with 95% confidence intervals based on robust standard errors. *First year* refers to the first year with available ifo survey responses of the firm and since survey data for the earliest decades is not available any more, it is *1980s* even when the firm started participation before that. *Last available year* refers to the decade of the last response in the survey and it is *2020s* for entities that still participate.

Shapley values (Shapley, 1953) which measure the individual marginal contribution to reach a common outcome.

Figure 3.9 shows the most important features as given by the SHAP[57] method. As is to be expected, name and address are the most relevant pieces of information. Furthermore, the Tfidf and SoftTfidf measures appear to be relatively important, whereas the name segments have a relatively smaller impact. While simple string similarity measures of name, street, and city contribute already much to the predicted probability, one can see that it is also important to incorporate different name variants such as previous names.

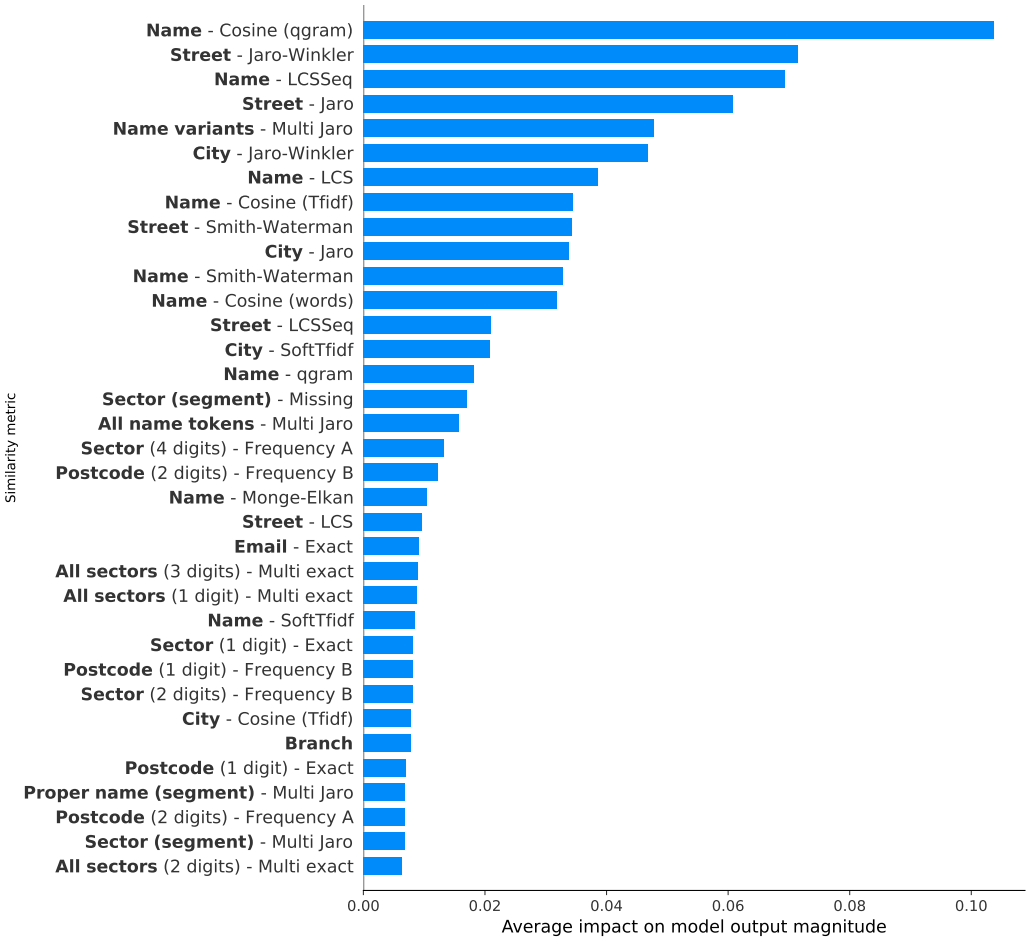
Selection Given the correlation of observed firm properties with the match rate, one may be concerned about how representative the matched sample is or about effects on downstream estimates[58]. Because the Business Panels are used for different types of research questions, it is not straightforward to test for selection bias in a general sense. Instead, in figure 3.10 I compare the time series of two of the most important questions[59]

[57] I used the official python implementation by the authors via the *shap* package.

[58] This is a general concern of RL applications and this topic has been the focus of several research papers such as Abowd and Vilhuber (2005), Moore et al. (2018), and Bailey et al. (2020).

[59] These two variables are for example used to compute the *ifo Business Cycle Index*.

Figure 3.9: SHAP feature importance



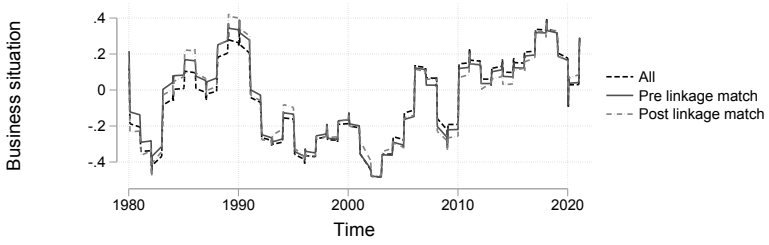
Note: Bars indicate the average absolute shapley values, i.e. the average impact of a similarity metric on the predicted probability. Only the 35 most relevant features out of 131 are presented here.

in the IBS, the assessment of the *business situation* (panel a) and the *business expectation* (panel b). The lines of both the pre and post linkages are close to the time series of all observations but there is nonetheless a difference which appears to decrease for the more recent periods.

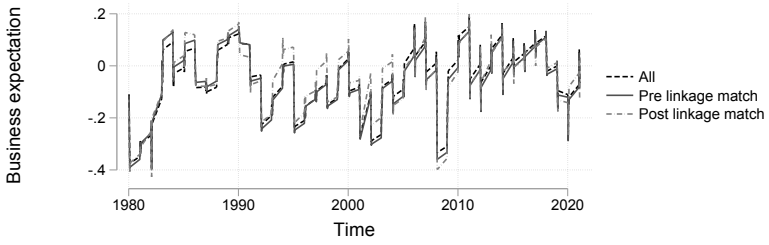
3.8 Discussion

While the linkage is overall successful, there is still room for improvement and possibilities for future linkages. One concern is computational efficiency: Implementing some parts programmed in pure Python in a more performant lower level language can speed them up. While this is sufficient for the volume of the two data sources, the method might not scale well to larger datasets. Therefore, some concepts of the literature focusing on this (e.g. Gschwind et al., 2019) can be applied in subsequent projects of the EBDC. Additionally, there are some techniques I could not fully utilize due to limitations of the protected PC used for data protection reasons. However, these methods, which include

Figure 3.10: Time series of business expectation and business situation by match status



(a) Business situation



(b) Business expectation

Note: The time series represent five months moving averages of the business situation and business expectation questions. The three level likert scale question has been recoded to 1 for *good*, 0 for *neutral*, and -1 for *bad*. The dark dashed line represents the time series for all firms, i.e., includes non-matched entities.

embedding and neural network based similarity metrics as well as segmentation of names in both data sources shall be explored in future iterations of this linkage.

Additionally, because some aspects of this linkage are specific to the present databases, is is not clear to what extent the method or trained models can be reused in other linkage applications with different data sources. However, it can be considered as a good starting point.

Preprocessing Because sector identifiers can be valuable in company linkage, it may be beneficial to further improve their cleaning and preparation. To this end, the full procedure from Link (2020) can help standardize the sectors better across surveys.[60] Another sector related concern is that the sector identifiers are time fixed such that a company that used to produce a product but eventually changed operations might not appear as a producing company in Orbis. This creates noise in the linkage and can lead to false negatives especially for survey respondents that no longer participate.[61]

Indexing There are a handful of indexing techniques that are more sophisticated than standard blocking and filtering and these can be applied to further increase the matchrate

[60] While the first digit identifier could be inferred for around 99% of ifo entities, a full 4 digit identifier is currently only available for 70% (as shown in table 3.2), thus creating more missings.

[61] This issue is less concerning if respondents continue to participate for example after a change in the main focus of production because according to Seiler and Heumann (2013), these would be considered to be a new entity.

and reduce the number of dissimilar pairs for faster linkage.[62] Another challenge comes with the *pre* and *post* linkages: While this paper tries to account for changes in relationships, it is difficult to encode this information into a linked research data set because the timing of the change is unknown. For this reason, the EBDC decided to only use the most recent match for each ID. In the future, one can try to find a way to systematically find a reliable date for the match change.

Comparison With the information gained from this paper, in future applications, the EBDC can reduce the number of features to the most relevant ones. This reduces the computation time of similarity metrics and further helps to avoid overfitting.

Classification To make the labeling of training data more efficient, I opted for an active learning approach where both cases predicted with high and low confidence were sampled. Because of this sampling procedure, the training data contain relatively more difficult cases and the evaluation metrics appear worse than they should be on a completely random sample. They can also not be compared to related linkages.

Another potential concern is that I used the same trained classifier ensemble for both the *pre* and *post* linkages even though there might be systematic differences. In a subsequent linkage, the pairs from the manual *post* corrections could be used as additional training data either in tandem with a dummy indicating the pre/post status or for a separate model.

A further challenge in RL is that it is not possible to compute similarity metrics for fields with missing information. Here, I assigned a field specific similarity of zero for pairs where a field is empty in either data source and additionally included a binary indicator for this value being missing. However, some of the classifiers that make up the ensemble, in particular the tree based models, can make use of this indicator better than others. Alternatively, one could use methods proposed in Ong et al. (2014) to handle these cases.

Postprocessing Despite extensive manual control, it is possible that there are still errors in the linkage given how complicated the task can be. Another challenge lies in choosing the correct BvD ID for each ifo ID if there are multiple predicted matches. Right now, I select the entity with the highest probability, irrespective of other matches. An alternative would be to only match this when it is sufficiently apart from a second potential match and to otherwise not match this at all. Because the key challenge is telling companies of the same corporate group apart, designing an optimization which takes similarity to other members of the group into account, as for example proposed in Mason (2018), may improve the linkage.

It might be further useful to incorporate information about mergers and acquisitions into the linkage to identify changes in associations and related entities. However, the EBDC currently does not have access to the separate *Zephyr* BvD database which contains such events.

[62] See Papadakis et al. (2019) for a survey on modern indexing methods.

3.9 Conclusion

Linked data offer great opportunities to work on novel research questions. However, linkage can be challenging, in particular when working with data from non-natural persons. The LMU-ifo EBDC offers researchers access to linked datasets which combine survey responses with financial information and wants to improve this offering by expanding the data as well as possible. Therefore, this paper combines the respondents of the ifo surveys to their respective records from the commercial Orbis database which contains financial information. Because there is no common identifier, I apply a probabilistic Record Linkage procedure supported by supervised ML for match classification. The process is tailored to the specific challenges of company data linkage via the use of appropriate similarity metrics and the exploration of NLP techniques.

The linkage works particularly well for more recent entries into the database where the entities have a very high match rate. Practically all false positives the classifier produced were cases where an entity was matched to a different but related company. This shows that the key difficulty in company RL is differentiating companies that are hierarchically related to one another.

Subsequent research should further explore the use of NLP techniques like for example Deep Learning based name similarity metrics which were not possible in the present application due to hardware limitations. Furthermore, because differentiating related companies from each other is the biggest challenge, it could be worthwhile to explore how information about the network for firms can be utilized. Ultimately, the majority of matches are oftentimes cases that are already very similar and a perfect probabilistic linkage will never be achieved such that some of the linkage techniques mostly serve to increase the match rate just a little bit more. Thus, an applied researcher must evaluate how much to invest in improvements into the linkage.

Appendices

Appendix A

Appendix to Chapter 1

A.1 Descriptives

Table A1: What medal is awarded to whom?

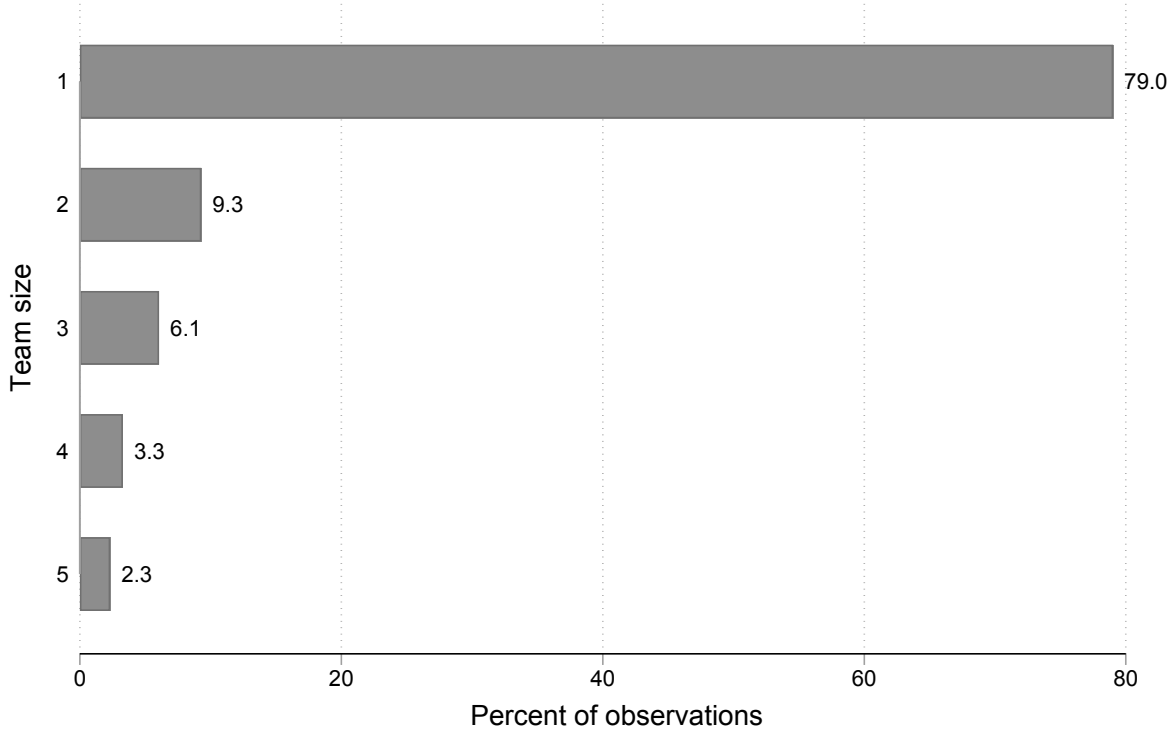
	Number of teams			
	0-99	100-249	250-999	1000+
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 +0.2%*	Top 10 + 0.2%*

* For every 500 further teams, an additional medal is awarded.

Table A2: Descriptive statistics on competitions

	Observations	Mean	SD	Min	Median	Max
Competitors	260	973.52	1347.54	4	393.50	9580
Teams	259	850.11	1184.29	6	351.00	8552
Submissions	260	18904.51	30190.32	0	5543.50	187626
Duration	259	81.85	52.34	3	76.00	731
Awards prize money	259	0.79	0.41	0	1.00	1
Prize money	248	43490.94	141170.56	0	20000.00	1500000
Medal winners	260	172.29	158.86	0	128.50	1152
Bronze medals	260	78.80	72.91	0	59.50	518
Silver medals	260	74.53	77.51	0	51.00	555
Gold medals	260	18.95	10.88	0	16.50	79
Money winners	260	5.58	5.08	0	4.00	41

Figure A1: Observations by team size



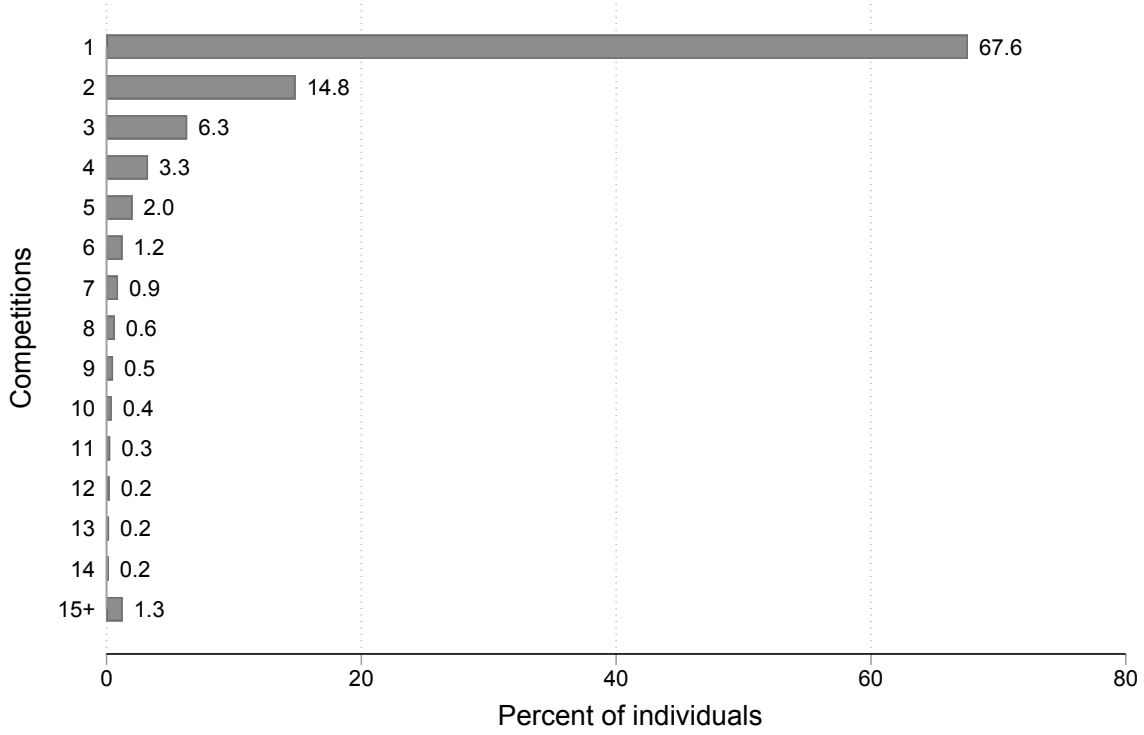
Note: Team size of 1 indicates solo participation. Teams are capped at five.

Table A3: Number of users by RDD and IV group assignment frequency

a) RDD:				
	Frequency in control			
Frequency in treatment	Never	1 time	2+ times	Total
Never	108609	4210	315	113134
1 time	4345	582	208	5135
2+ times	472	200	244	916
Total	113426	4992	767	119185
b) IV:				
Frequency in treatment	Never	1 time	2+ times	Total
Never	105717	5182	496	111395
1 time	5539	746	290	6575
2+ times	626	306	283	1215
Total	111882	6234	1069	119185

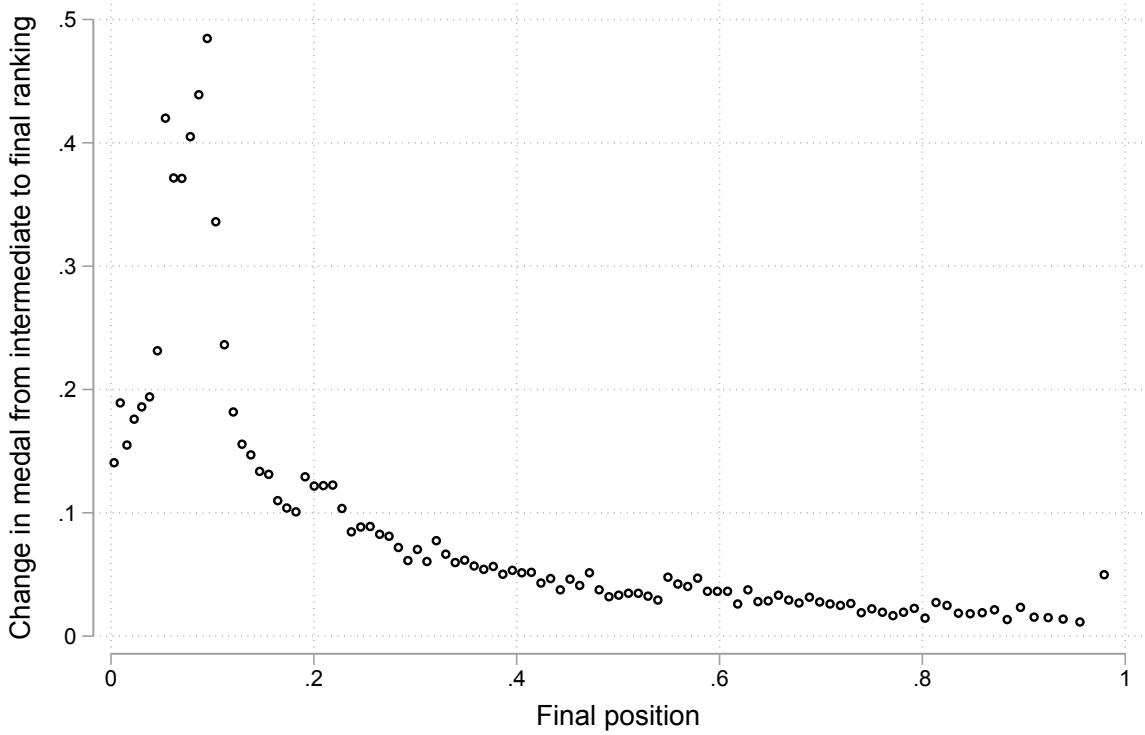
Note: For each competition, users are assigned to treatment or control group when their final rank is respectively to the right and left of the running variable within a window of two and a half percent of the leaderboard.

Figure A2: Number of competitions



Note: Individuals that participate 15 times or more are pooled in the bottom bar.

Figure A3: Medal changes by final leaderboard position



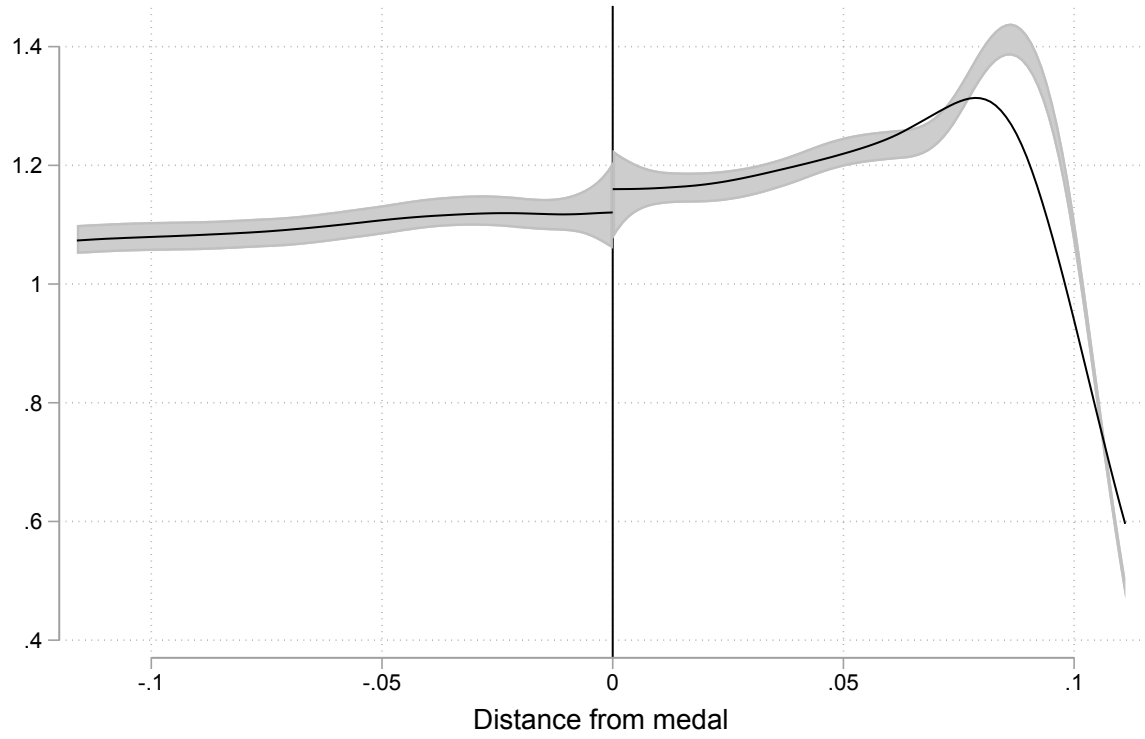
Note: Binned scatterplot showing the share of participants who won or lost a medal due to a rank change by position on the final leaderboard.

Table A4: Balancing of covariates by distance from medal

	At-competition values	Pre-competition means
Private score		0.01137
		(0.01448)
Public score		0.00977
		(0.01474)
Intermediate position		-0.01808
		(0.01478)
Score change	-0.00840**	0.00180
	(0.00404)	(0.00466)
Rank change percentile	-0.00006	-0.00003*
	(0.00005)	(0.00002)
Competes in team	-0.01443	0.03259
	(0.02541)	(0.02335)
Team size	-0.05446	0.04962
	(0.07330)	(0.05084)
New team	-0.03086	0.02981
	(0.02450)	(0.02194)
Submissions	-4.27082*	1.56924
	(2.43075)	(2.02795)
Used public code	0.00528	0.00125
	(0.00411)	(0.00577)
Published own code	-0.00172	0.00142
	(0.00352)	(0.00321)
Experience (competitions)	0.10377	
	(0.50987)	
N previous bronze	-0.05900	
	(0.09845)	
N previous silver	0.04286	
	(0.10205)	
N previous gold	0.01511	
	(0.05549)	
Cum. money wins	-0.00442	
	(0.01780)	

Note: Standard errors in parenthesis. Standard errors are clustered on team-competition level in the left column and robust standard errors in the right column. * p<0.1, ** p<0.05, *** p<0.01. Values represent Regression Discontinuity estimates. Left column: balancing of variables at the time of the treatment competition. I.e. how do observations to the left and right of the threshold differ? Right column: balancing of variables before the treatment competition, aggregated on a user level. I.e.: how do individuals to the left and right of the threshold differ with respect to their history? For some variables, balancing is only informative in either the left or right column, and they are omitted otherwise.

Figure A4: Continuity of the running variable: Relative distance to bronze rank



Note: The graph shows the distribution of the running variable, the distance from a bronze medal, around the medal winning threshold. A window of around 10% of the leaderboard on either side is shown here. The gray area represents the 95% confidence band.

A.2 Estimation details

A.2.1 Comparison of the two identification approaches

Both identification strategies estimate different Local Average Treatment Effects (LATEs) and have different strengths. The RDD has a simple setup that is straightforward and intuitive to interpret, while the IV requires better knowledge about the competition procedure. Related to this is that the IV requires to make additional assumptions for identification of a causal effect, namely relevance and exogeneity of the instrument. The critical assumption of the RDD, namely that participants cannot influence their exact placement, is trivially given in a competition setting where one cannot influence the placement of other competitors. At the same time, the IV also has some advantages: In particular, given a sufficiently large score change, the effect is also identified from silver and gold medals. The RDD can instead only measure the effect of a bronze medal relative to none, silver relative to bronze, or gold relative to silver. However, this makes the RDD more intuitive to interpret. The RDD is also very narrow in terms of the skill level of individuals that contribute to the LATE, whereas the compliers of the IV are potentially less homogeneous.

A.2.2 Results

Table A5: IV estimation results: Effect on team formation

a) First stage:			
	(1)	(2)	(3)
	Medal winner (t-1)	Medal winner (t-1)	Medal winner (t-1)
Score change (t-1)	0.90113***	0.90113***	0.88826***
	(0.02239)	(0.02239)	(0.02355)
b) Second stage:			
	(1)	(2)	(3)
	Competes in team	New team	Switch to team
Medal winner (t-1)	0.05135**	0.00329	0.02154
	(0.02372)	(0.02135)	(0.02375)
Observations	100602	100602	85181
Individuals	28393	28393	24671
1st stage F	1620.09	1620.09	1422.27

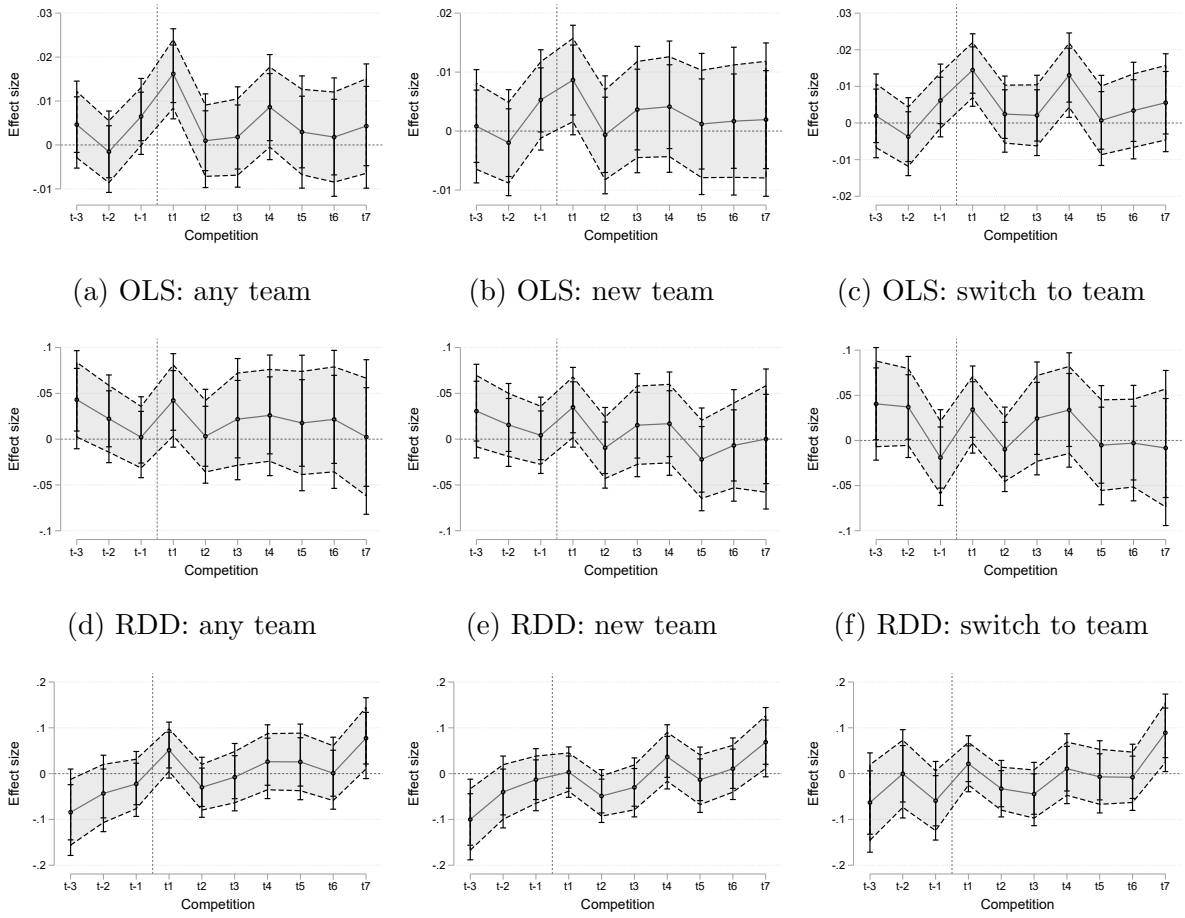
Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. For the first stage, estimates for other covariates are not reported in this table Includes fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A6: Estimation results - Effect on signaling - IV

a) First stage:					
	(1)	(2)	(3)	(4)	(5)
	Medal winner	Medal winner	Medal winner	Medal winner	Medal winner
Score change	1.63665***	1.70882***	1.70882***	1.70882***	1.70882***
	(0.04082)	(0.16838)	(0.16838)	(0.16838)	(0.16838)
b) Second stage:					
	(1)	(2)	(3)	(4)	(5)
	Links LinkedIn	Mention competitions	Certificates	Recommendations	Skills
Medal winner	0.07036***	0.22325**	0.09979	0.17832	-1.15382
	(0.02321)	(0.10597)	(1.24861)	(0.63062)	(2.70891)
Observations	9445	1103	1103	1103	1103
Individuals	8637	935	935	935	935
1st stage F	1607.63	102.99	102.99	102.99	102.99

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Estimates for other covariates are excluded for the first stage. Includes competition fixed effects.

Figure A5: Coefficients for different periods between winning the medal and the outcome



(a) OLS: any team (b) OLS: new team (c) OLS: switch to team

(d) RDD: any team (e) RDD: new team (f) RDD: switch to team

(g) IV: any team (h) IV: new team (i) IV: switch to team

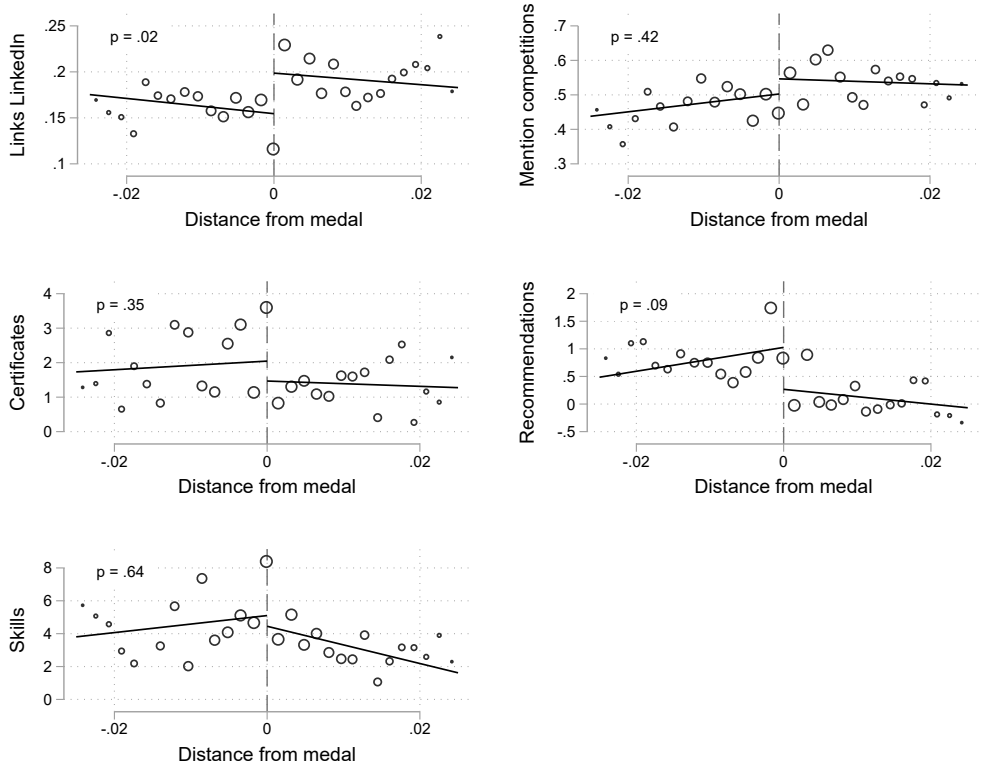
Note: Gray areas represent 95% confidence intervals and solid lines indicate both 90% and 99% confidence intervals. Clustered at team-competition level. The horizontal dashed line is at zero. The vertical dashed line separates outcome competitions before the treatment competition from those after the treatment competition. The outcome competition at timing c_0 is omitted. All estimates result from separate regressions. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. Coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A7: Estimation results - Effect on labor market success - IV

a) First stage:	
	(1)
	Medal winner
Score change	1.70590***
	(0.16776)
b) Second stage:	
	(1)
	Data Scientist
Medal winner	0.03959
	(0.10832)
Observations	1103
Individuals	935
1st stage F	103.40

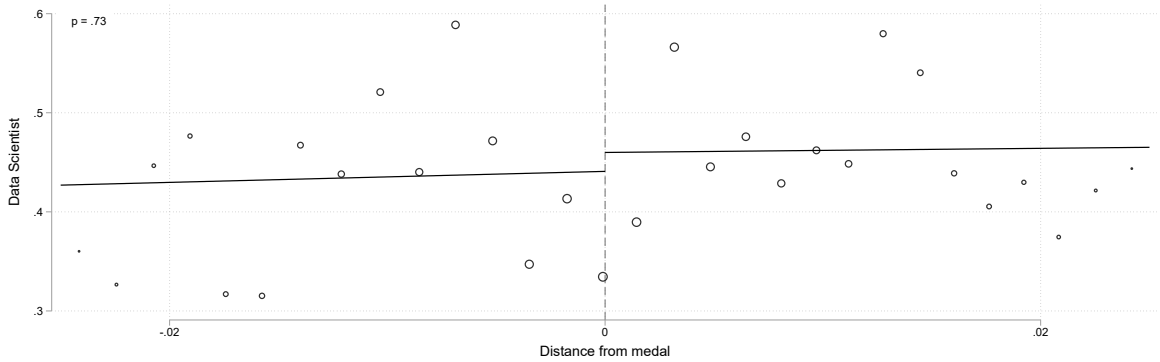
Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. Estimates for other covariates are excluded for the first stage. Includes competition fixed effects.

Figure A6: RDD plots - Effects on signaling activity



Note: Binned scatterplots with local linear regression. The p-value of the discontinuity estimate is displayed in the top left corner. The size of the circles shows the relative average weight of observations within each bin. Weights are given by the triangular kernel function and decrease linearly with the distance from the cutoff. Only the area of 2.5% of the leaderboard on both sides of the cutoff are taken into account for the regression and displayed here. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Includes competition fixed effects.

Figure A7: RDD plots - Effects on labor market success



Note: Binned scatterplots with local linear regression. The p-value of the discontinuity estimate is displayed in the top left corner. The size of the circles shows the relative average weight of observations within each bin. Weights are given by the triangular kernel function and decrease linearly with the distance from the cutoff. Only the area of 2.5% of the leaderboard on both sides of the cutoff are taken into account for the regression and displayed here. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. Includes competition fixed effects.

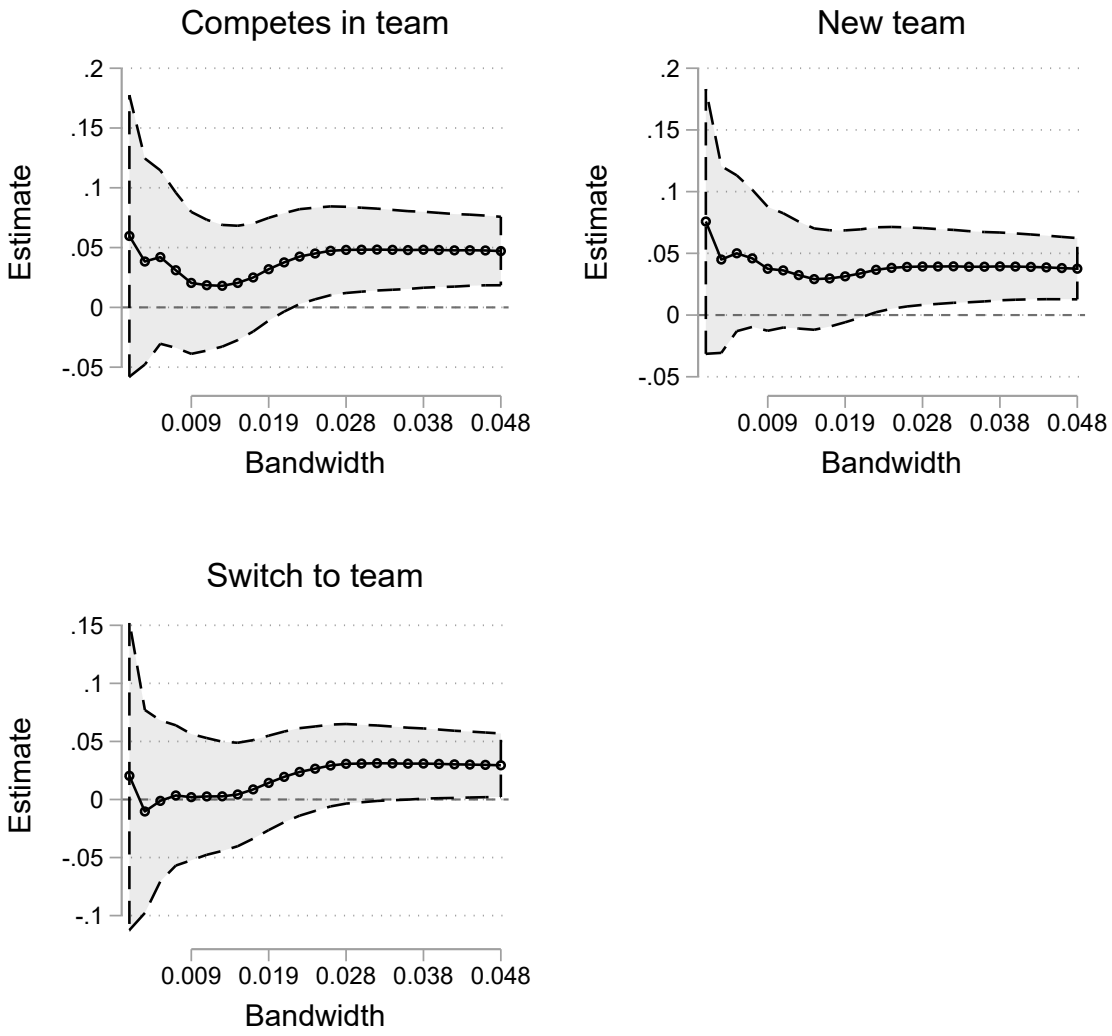
A.3 Alternative specifications

Table A8: Estimation results: Effects on team formation - OLS without individual fixed effects

	(1) Competes in team	(2) New team	(3) Switch to team
Medal winner (t-1)	0.04259***	0.03079***	0.03693***
	(0.00385)	(0.00327)	(0.00367)
Observations	106734	106734	89633
Individuals	29296	29296	25381

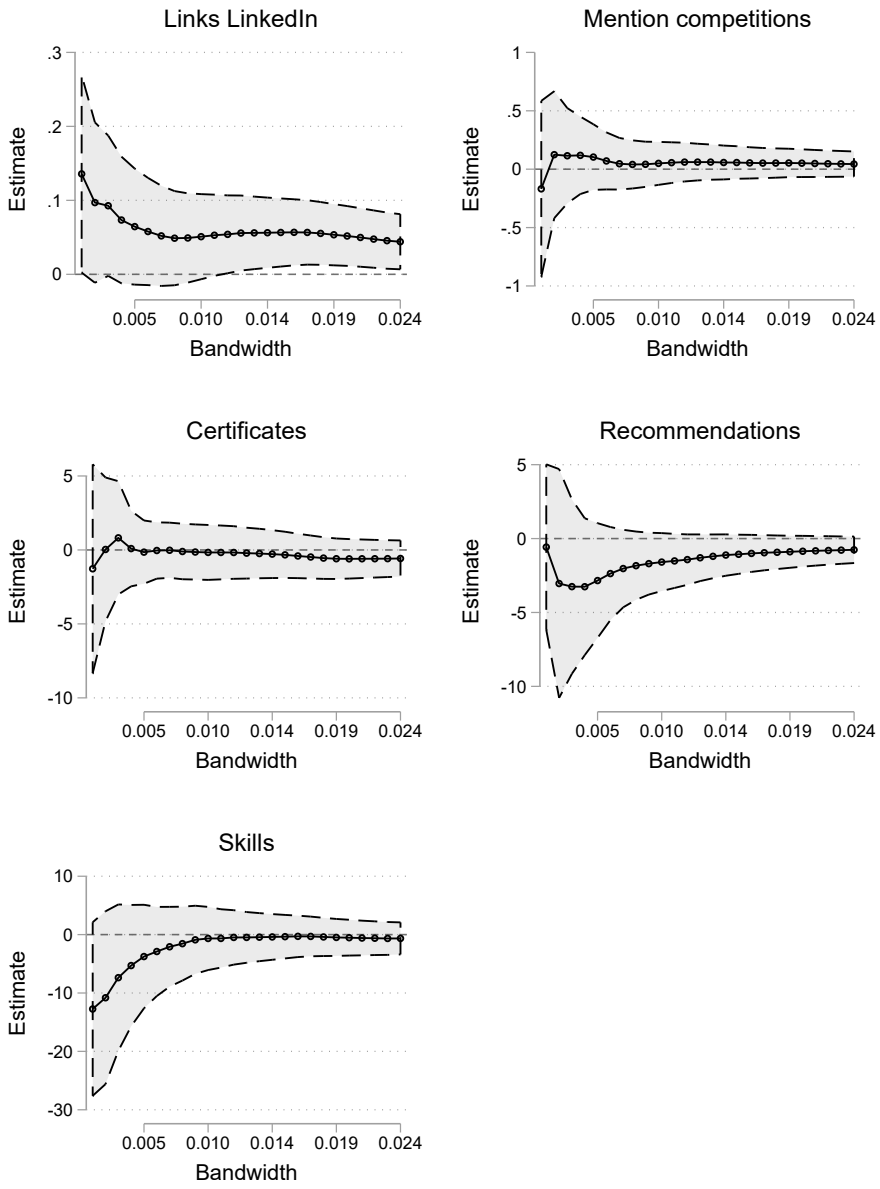
Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. Estimates for other covariates are excluded. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. Includes fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Figure A8: Different bandwidthes: Effects of team formation



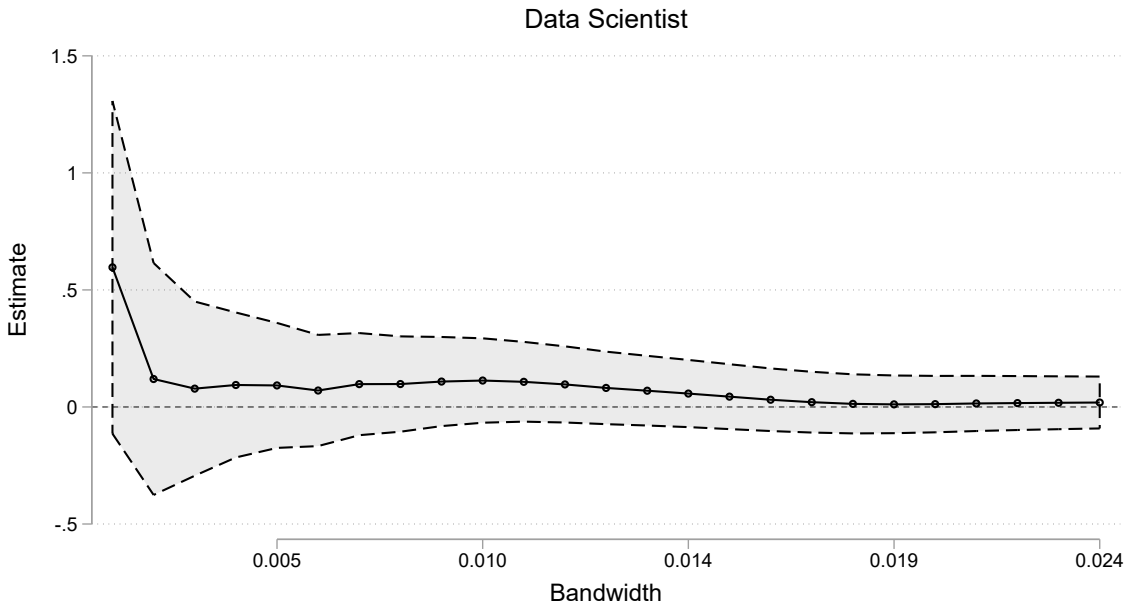
Note: RDD treatment effects for different bandwidthes. The dots show the point estimates and the gray area represents the 95% confidence bands. For all bandwidthes, a first order polynomial was fit with triangular kernel and the full set of covariates is included. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. Includes competition fixed effects.

Figure A9: Different bandwidthes: Signaling activity



Note: RDD treatment effects for different bandwidthes. The dots show the point estimates and the gray area represents the 95% confidence bands. For all bandwidthes, a first order polynomial was fit with triangular kernel and the full set of covariates is included. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Includes competition fixed effects.

Figure A10: Different bandwidthes: Labor market success



Note: RDD treatment effects for different bandwidthes. The dots show the point estimates and the gray area represents the 95% confidence bands. For all bandwidthes, a first order polynomial was fit with triangular kernel and the full set of covariates is included. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. Includes competition fixed effects.

Table A9: Estimation results: Effects on labor market success - no controls for unrelated signals

	(1) Data Scientist
a) OLS:	
Medal winner	0.01756**
	(0.00853)
Observations	22493
Individuals	2859
b) RDD:	
Medal winner	0.01234
	(0.05642)
Observations	1686
Individuals	1450
c) IV:	
Medal winner	0.04228
	(0.10830)
Observations	1103
Individuals	935
1st stage F	102.99

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

Table A10: Effect of incrementally better medals on team formation

	(1) Competes in team	(2) New team	(3) Switch to team
a) RDD Silver:			
Silver (t-1)	-0.01187	-0.01136	-0.01662
	(0.01961)	(0.01592)	(0.01857)
Observations	8696	8696	6268
Individuals	4096	4096	2989
b) RDD Gold:			
Gold (t-1)	-0.01054	0.00280	-0.00338
	(0.02118)	(0.01848)	(0.02950)
Observations	8542	8542	4094
Individuals	3093	3093	1797
c) IV Silver:			
Silver (t-1)	0.06705**	0.00429	0.03131
	(0.03098)	(0.02788)	(0.03453)
Observations	100602	100602	85181
Individuals	28393	28393	24671
1st stage F	1442.28	1442.28	1122.95
d) IV Gold:			
Gold (t-1)	0.19446**	0.01245	0.13823
	(0.09012)	(0.08086)	(0.15261)
Observations	100602	100602	85181
Individuals	28393	28393	24671
1st stage F	586.85	586.85	318.64

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A11: Effect of incrementally better medals on team formation

	Comp. signals		Oth. signals		
	(1) Links LinkedIn	(2) Mention competitions	(3) Certificates	(4) Recommendations	(5) Skills
a) IV Silver:					
Silver	0.11867***	0.32501**	0.14528	0.25961	-1.67979
	(0.03923)	(0.15430)	(1.81651)	(0.91820)	(3.95502)
Observations	9445	1103	1103	1103	1103
Individuals	8637	935	935	935	935
1st stage F	951.14	92.91	92.91	92.91	92.91
b) IV Gold:					
Gold	0.62805***	2.45469*	1.09728	1.96076	-1.3e+01
	(0.20927)	(1.27282)	(13.72124)	(6.94401)	(29.88697)
Observations	9445	1103	1103	1103	1103
Individuals	8637	935	935	935	935
1st stage F	331.37	17.13	17.13	17.13	17.13

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A12: Effect of incrementally better medals on team formation

	(1) Data Scientist
a) IV Silver:	
Silver	0.05754
	(0.15739)
Observations	1103
Individuals	935
1st stage F	92.62
b) IV Gold:	
Gold	0.43697
	(1.19976)
Observations	1103
Individuals	935
1st stage F	16.92

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A13: Estimation results: Effects on team formation - excludes teams that are the same as in t-1

	(1) Competes in team	(2) New team	(3) Switch to team
a) OLS:			
Medal winner (t-1)	0.01563***	0.00865**	0.01446***
	(0.00394)	(0.00361)	(0.00384)
Observations	92022	93468	78153
Individuals	28914	29296	25381
b) RDD:			
Medal winner (t-1)	0.03574*	0.03482**	0.03421*
	(0.01927)	(0.01690)	(0.01871)
Observations	7506	7648	6163
Individuals	3700	3768	2978
c) IV:			
Medal winner (t-1)	0.03729	0.00329	0.02154
	(0.02345)	(0.02135)	(0.02375)
Observations	99138	100602	85181
Individuals	28023	28393	24671
1st stage F	1592.40	1620.09	1422.27

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. Only users that have ever competed in a team before are included. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A14: Estimation results: Effects on team formation - no pure solo participants

	(1) Competes in team	(2) New team	(3) Switch to team
a) OLS:			
Medal winner (t-1)	0.02122***	0.01038*	0.02019**
	(0.00700)	(0.00620)	(0.00868)
Observations	36302	36302	21422
Individuals	9402	9402	4870
b) RDD:			
Medal winner (t-1)	0.08106**	0.06118**	0.05095
	(0.03685)	(0.02917)	(0.04037)
Observations	3428	3428	1939
Individuals	1636	1636	823
c) IV:			
Medal winner (t-1)	0.11313**	-0.01712	0.06101
	(0.05205)	(0.04432)	(0.07237)
Observations	37043	37043	21612
Individuals	9006	9006	4705
1st stage F	456.09	456.09	222.84

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. Only users that have ever competed in a team before are included. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A15: Estimation results: Effects on signaling - one-time-compliers

	Comp. signals		Oth. signals		
	(1) Links LinkedIn	(2) Mention competitions	(3) Certificates	(4) Recommendations	(5) Skills
a) RDD one-time-complier:					
Medal winner	0.03607*	0.03459	-0.55858	-0.66500	0.10557
	(0.01955)	(0.06376)	(0.72296)	(0.45797)	(1.53057)
Observations	8027	1180	1180	1180	1180
Individuals	8037	1229	1229	1229	1229
b) IV one-time-complier:					
Medal winner	0.04963**	0.23178	-0.31055	0.40411	-0.03977
	(0.02358)	(0.14246)	(1.84365)	(0.96880)	(3.72045)
Observations	7615	682	682	682	682
Individuals	7646	715	715	715	715
1st stage F	1366.22	57.46	57.46	57.46	57.46

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects. The *one-time-compliers* includes only individuals that are only a single time in either control or treatment group for the RDD and those that are only a single time in the group of compliers for the IV

Table A16: Estimation results: Effects on labor market success - one-time-compliers

	(1) Data Scientist
a) RDD one-time-complier:	
Medal winner	0.06554
	(0.06566)
Observations	1180
Individuals	1229
b) IV one-time-complier:	
Medal winner	0.01078
	(0.14357)
Observations	682
Individuals	715
1st stage F	57.66

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. All specifications include competition fixed effects. The *one-time-compliers* includes only individuals that are only a single time in either control or treatment group for the RDD and those that are only a single time in the group of compliers for the IV

Table A17: Estimation results: Effects on signaling with controls for later medals

	Comp. signals		Oth. signals		
	(1) Links LinkedIn	(2) Mention competitions	(3) Certificates	(4) Recommendations	(5) Skills
a) OLS:					
Medal winner	0.01493***	0.07901***	-0.83374***	-0.40083***	-0.46237**
	(0.00307)	(0.00706)	(0.11933)	(0.04982)	(0.20589)
Observations	158331	22493	22493	22493	22493
Individuals	53371	2859	2859	2859	2859
b) RDD:					
Medal winner	0.03368*	0.00063	-0.69258	-0.68448	-0.50070
	(0.01889)	(0.05430)	(0.62909)	(0.45584)	(1.43646)
Observations	9709	1686	1686	1686	1686
Individuals	8790	1450	1450	1450	1450
c) IV:					
Medal winner	0.05971***	0.18074*	0.08299	0.23231	-0.85488
	(0.02299)	(0.10384)	(1.26187)	(0.63666)	(2.73742)
Observations	9445	1103	1103	1103	1103
Individuals	8637	935	935	935	935
1st stage F	1599.92	100.75	100.75	100.75	100.75

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins respectively won before and after the competition, used public code, and published own code. All specifications include competition fixed effects.

Table A18: Estimation results: Effects on labor market success with controls for later medals

	(1) Data Scientist
a) OLS:	
Medal winner	0.01879**
	(0.00854)
Observations	22493
Individuals	2859
b) RDD:	
Medal winner	-0.00593
	(0.05644)
Observations	1686
Individuals	1450
c) IV:	
Medal winner	0.02239
	(0.10871)
Observations	1103
Individuals	935
1st stage F	101.26

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins respectively won before and after the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. All specifications include competition fixed effects.

Table A19: Estimation results: Effects on continuing participation

	(1) OLS	(2) RDD	(3) IV
Medal winner	0.03085***	0.00174	0.21396***
	(0.00268)	(0.01646)	(0.02711)
Observations	169847	10576	229789
RDD Bandwidth	-	+/- 0.025	-
IV 1st stage F	-	-	2778.97

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

Table A20: Estimation results: Effects on team formation - Heckman correction

	(1) Competes in team	(2) New team	(3) Switch to team
a) OLS:			
Medal winner (t-1)	0.01114	0.00578	0.02233**
	(0.00747)	(0.00682)	(0.00871)
Observations	80183	80183	66359
Individuals	19066	19066	16306
b) RDD:			
Medal winner (t-1)	0.01664	0.00495	0.01028
	(0.02033)	(0.01743)	(0.01921)
Observations	7347	7347	5912
Individuals	3614	3614	2849
c) IV:			
Medal winner (t-1)	0.03139	-0.00907	0.00274
	(0.02548)	(0.02285)	(0.02503)
Observations	81776	81776	68456
Individuals	18440	18440	15831
1st stage F	1403.60	1403.60	1261.36

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. Selection has been modeled as a function of winning a medal *linked LinkedIn*, *linked Github*, *linked own URL*, the number of followers, and other covariates. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. OLS includes individual fixed effects. All specifications include fixed effects for competitions at time of the outcome and competitions at time of the treatment.

Table A21: Estimation results: Effects on signaling - Heckman correction

	Comp. signals	Oth. signals		
	(1) Mention competitions	(2) Certificates	(3) Recommendations	(4) Skills
a) OLS:				
Medal winner	0.08917***	-0.91392***	-0.38209***	-0.48867**
	(0.00715)	(0.11954)	(0.05099)	(0.20608)
Observations	22493	22493	22493	22493
Individuals	2859	2859	2859	2859
b) RDD:				
Medal winner	0.04358	-0.63718	-0.76681*	-0.63685
	(0.05504)	(0.62967)	(0.44746)	(1.39815)
Observations	1686	1686	1686	1686
Individuals	1450	1450	1450	1450
c) IV:				
Medal winner	0.22264**	0.10810	0.18079	-1.14048
	(0.10573)	(1.24568)	(0.62958)	(2.70382)
Observations	1103	1103	1103	1103
Individuals	935	935	935	935
1st stage F	105.02	105.02	105.02	105.02

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. Selection has been modeled as a function of winning a medal *linked Github*, *linked own URL*, the number of followers, and other covariates. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

Table A22: Estimation results: Effects on labor market success - Heckman correction

	(1) Data Scientist
a) OLS:	
Medal winner	0.02243***
	(0.00852)
Observations	22493
Individuals	2859
b) RDD:	
Medal winner	0.01870
	(0.05653)
Observations	1686
Individuals	1450
c) IV:	
Medal winner	0.03945
	(0.10816)
Observations	1103
Individuals	935
1st stage F	105.66

Note: Standard errors in parenthesis, clustered on team-competition level; * p<0.1, ** p<0.05, *** p<0.01. For the binary outcomes, coefficients of the linear probability models can be interpreted as percentage point changes in the outcomes. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. Selection has been modeled as a function of winning a medal *linked Github*, *linked own URL*, the number of followers, and other covariates. RDD: Polynomial of degree one, i.e. a linear curve, fit to both sides of the cutoff. Observations are weighted with a triangular kernel function. IV: Only second stage reported. Estimates for other covariates are excluded. The covariates are the same for all specifications and include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. All specifications include competition fixed effects.

Table A23: Alternative RDD specifications - Team formation

	(1) Competes in team	(2) New team	(3) Switch to team
Kernel=triangular	0.04228**	0.03482**	0.03421*
	(0.01980)	(0.01690)	(0.01871)
Kernel=uniform	0.04744**	0.04064**	0.03825**
	(0.01915)	(0.01616)	(0.01806)
Kernel=epanechnikov	0.04656**	0.03670**	0.03802**
	(0.01952)	(0.01668)	(0.01851)

Note: Each line represents Regression Discontinuity estimates for a different set of polynomial fit and kernel function. *Poly* refers to the degree of the polynomial fit on each side of the cutoff. All specifications include covariates. *Competes in team* is a dummy indicating team rather than solo participation in competition at time t_0. *New team* is a dummy indicating that the individual participated in a team in the competition at time t_0, where they have never participated with any of the team members before. *Switch to team* is a dummy indicating that the individual participated solo in the competition at time t_0 and in a team with others at time t_1. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins, used public code, and published own code. All specifications include competition fixed effects.

Table A24: Alternative RDD specifications - Signaling activity

	(1) Links LinkedIn	(2) Mention competitions	(3) Certificates	(4) Recommendations	(5) Skills
Kernel=triangular	0.04410**	0.04392	-0.57776	-0.75949*	-0.65079
	(0.01900)	(0.05502)	(0.62166)	(0.45605)	(1.40089)
Kernel=uniform	0.03313*	0.03562	-0.53317	-0.56040*	-0.46966
	(0.01764)	(0.05055)	(0.57460)	(0.33974)	(1.26198)
Kernel=epanechnikov	0.04075**	0.04067	-0.57964	-0.68019*	-0.69019
	(0.01854)	(0.05350)	(0.60912)	(0.41144)	(1.34835)

Note: Each line represents Regression Discontinuity estimates for a different set of polynomial fit and kernel function. *Poly* refers to the degree of the polynomial fit on each side of the cutoff. All specifications include covariates. *Links LinkedIn* is a dummy indicating that the individual provided a link to their LinkedIn profile on their competition profile. *Mention competitions* is a dummy indicating that the individual mentions the competition website on their LinkedIn profile. *Certificates* and *Skills* are the number of certificates and skills respectively publicly listed on an individual's LinkedIn profile. *Recommendations* is the number of recommendations by other people listed publicly on an individual's LinkedIn profile. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. All specifications include competition fixed effects.

Table A25: Alternative RDD specifications - Labor market success

	(1) Data Scientist
Kernel=triangular	0.01909 (0.05651)
Kernel=uniform	0.01894 (0.05325)
Kernel=epanechnikov	0.01841 (0.05567)

Note: Each line represents Regression Discontinuity estimates for a different set of polynomial fit and kernel function. *Poly* refers to the degree of the polynomial fit on each side of the cutoff. All specifications include covariates. *Data Scientist* is a dummy indicating that the individual works as a Data Scientist or in a related profession. The covariates include experience in terms of competitions, team size dummies, number of submissions, cumulative sum of bronze, silver, gold medals, and prize money wins won before the competition, used public code, and published own code. Additional covariates included for signaling behavior: links LinkedIn, reffering competitions, certificates, recommendations, and skills listed on the résumé. All specifications include competition fixed effects.

Appendix B

Appendix to Chapter 2

B.1 Omitted figures

B.1.1 Exemplary comments

The purpose of a publicly traded company is to generate a sufficient rate of return. And it's every investor's right to exert pressure on the company management. Lamentations out of place.

→ *work* → *money*

That's right. I forgot about family reunion. Wouldn't have thought that young men leave their women and kids to come to Germany.

→ *family*

It's not about proving something. Noone should be forced to join a demonstration. I guess I wouldn't have gone myself, because I'm a lazy bastard. But it's a scandal that official associations distance themselves from demos.

Figure B1: Topic classification: three examples

Notes: Comment 1 is classified as *work* and *money*, comment 2 is classified as *family*, and comment 3 is classified as not covering a gender stereotypical topic.

B.1.2 Further indices

I have been saying for long what Mr. Steinbrück said. It's a pity that he was so abandoned by his party allies during the election campaign, especially by Mr. Gabriel. A true Social Democrat who has been impressed and fostered by Helmut Schmidt. Back then, he was the only MP who would disclose his income. The others were too craven and mocked him. He is the most sincere politician whom I know. I can only beg him to return to policy and to show and teach his party allies true Social Democratic policy.

→ *male*

Ms. Kässmann is and will be an idol to me. Smart, good-looking, courageous, warm, coherent, good mother, faithful Christian. The Protestant Church did not suffer, of course, to the contrary. People set standards for dealing with fault.

→ *female*

The author did not care about whether the small sales are truly just due to the design or due to the price as well. I personally prefer to drive a rare car on German streets, a Daihatsu-Copen... Even after 6 years people keep asking me what kind of car that is. However, the Copen is too small for many people, because it just has 2 sears and a small trunk, where in the summer the roof is stored to drive overtly. I also have to say that the Copen was only available as right-hand drive car in its first years. Tuning pieces are only available in Japan for high prices, plus German customs with more than 20%.

→ *none*

Figure B2: Gender classification: three examples

Notes: Comment 1 is classified as *male*, comment 2 is classified as *female*, and comment 3 is classified as neither *male* nor *female*.

Figure B3: Sentiment-weighted indices

Notes: The figure displays our sentiment-weighted indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

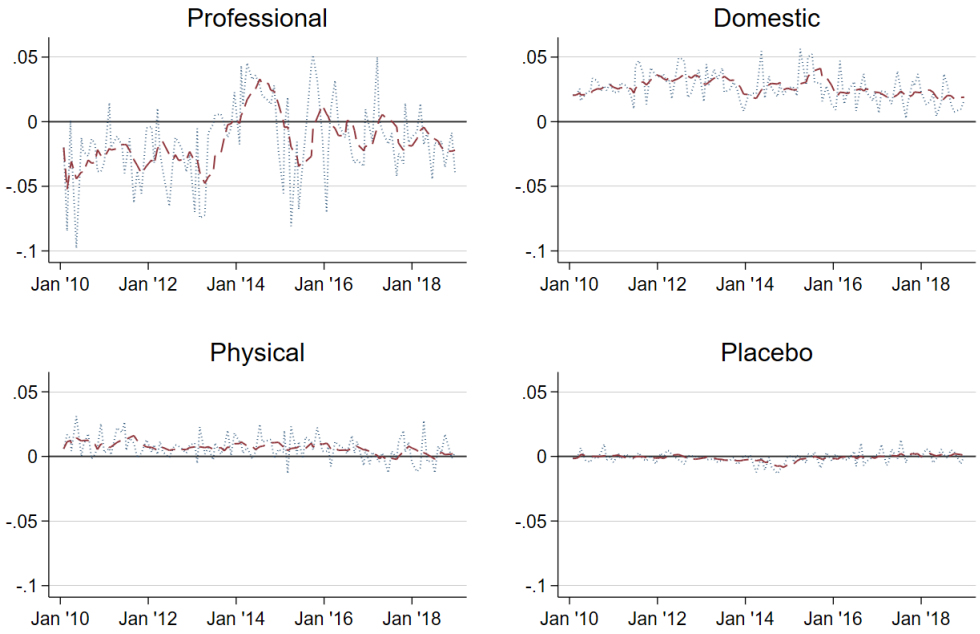


Figure B4: Exclude comments with offensive language

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on a subsample that excludes all comments classified as *offensive*. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

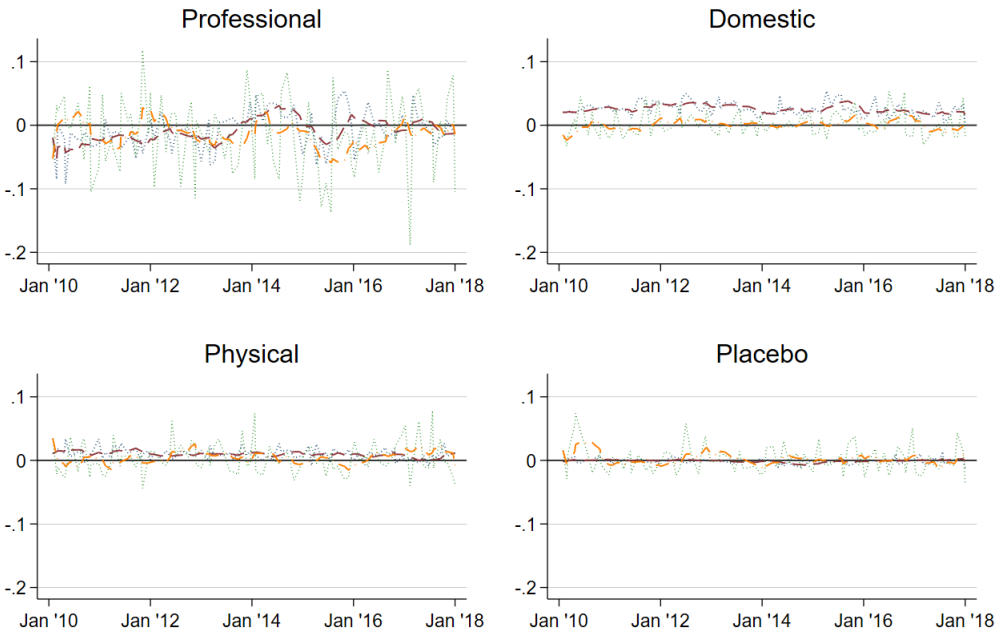


Figure B5: Gender stereotypes in news articles

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The blue and the green dotted lines correspond to the index as illustrated in Section 2.4.2.1, where the blue line is based on comments, and the green line is based on news articles. The red and orange dashed lines correspond to a moving average based on the current and the five previous months, where the red line is based on the indices for comments, and the red line based on the indices for news articles.

Figure B6: Indices based on residuals

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*, based on the residuals from an OLS regression of each topic indicator on observable comment and user characteristics. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

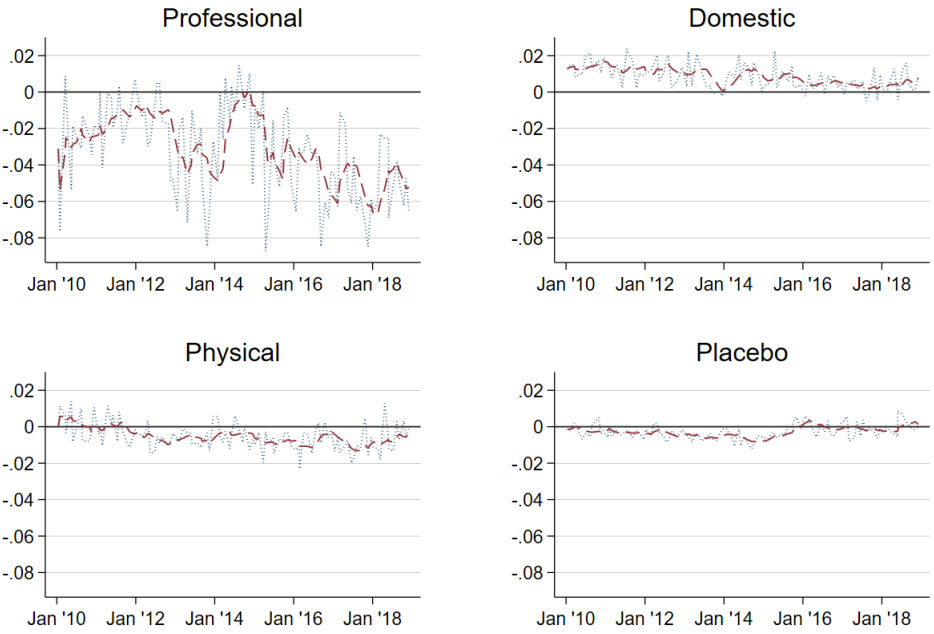


Figure B7: Include comments on Angela Merkel

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on a sample that includes all comments on Angela Merkel. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

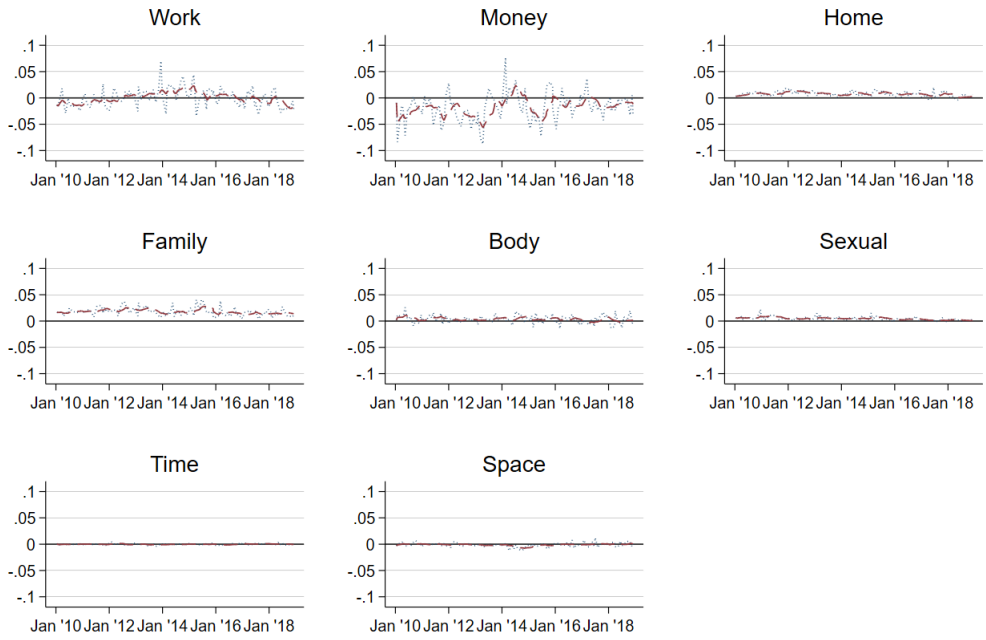


Figure B8: Non-pooled topics

Notes: The figure displays our indices for the topics *work*, *money*, *home*, *family*, *body*, *sexual*, *time*, and *space*. In contrast to our main analysis, related topics are not pooled together. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

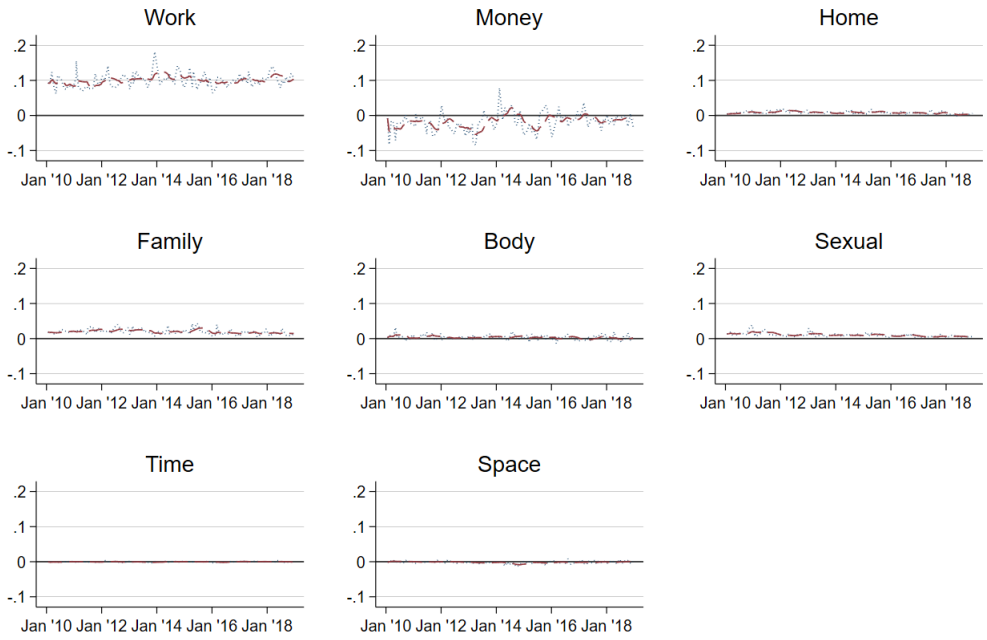


Figure B9: Non-binary classification

Notes: The figure displays our indices for the topics *work*, *money*, *home*, *family*, *body*, *sexual*, *time*, and *space*. In contrast to our main analysis, related topics are not pooled together. Moreover, the index is based on raw continuous probabilities for topics instead of topic indicators. The dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

Figure B10: Alternative gender classification

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on an alternative gender classification as illustrated in Section 2.5. Otherwise, the dotted line corresponds to the index as illustrated in Section 2.4.2.1. The dashed line corresponds to a moving average based on the current and the five previous months.

B.2 Omitted tables

Table B1: Wikipedia categories

LIWC topic	Wikipedia category	no. articles
work	working environment	62
money	finance	48
	means of payment	28
	money transfers	143
family	family	72
	family model	8
	relatives (male)	16
	relatives (female)	4
	terms of relativeness	14
home	housekeeping	50
	types of rooms	151
	apartment	25
	apartment (part of building)	16
body	physiques	27
	body extent	67
	body region	12
sexual	human sexuality	100

Notes: Table B1 shows the Wikipedia categories, the corresponding gender stereotypical topics, and the associated number of Wikipedia articles that we retrieve to validate our topic classification procedure.

Technical details

This section provides some technical details on our *clustered tf-idf* approach.

Notation Suppose that there D documents (here: comments) and J clusters of words with similar meaning. Each cluster j comprises W_j words, where W_j is small. Denote these words by $w_{1,j}, ..., w_{k,j}, ..., w_{W_j,j}$.

Clustered term frequency (ctf) The clustered term frequency for cluster $j \in \{1, ..., J\}$ and document $d \in \{1, ..., D\}$ is given by

$$\text{ctf}_{j,d} = \frac{\sum_{k=1}^{W_j} \#(w_{k,j} \text{ in } d)}{\max\left(\sum_{w=1}^{W_1} \#(w_{k,1} \text{ in } d), \sum_{w=1}^{W_2} \#(w_{k,2} \text{ in } d), ..., \sum_{w=1}^{W_J} \#(w_{k,J} \text{ in } d)\right)}, \tag{B.1}$$

where $\#(w_{k,j} \text{ in } d)$ is the number of occurrences of word $w_{k,j}$ in document d.

Clustered inverse document frequency (cidf) The clustered inverse document frequency for cluster $j \in \{1, ..., J\}$ and document $d \in \{1, ..., D\}$ is given by

$$\text{cidf}_{j,d} = \log\left(\frac{N}{\max\left(\text{df}(w_{1,j}), \text{df}(w_{2,j}), ..., \text{df}(w_{W_j,j})\right) + 1}\right), \tag{B.2}$$

where $\text{df}(w_{k,j})$ is the document frequency of word $w_{k,j}$.

Clustered tf-idf Given the clustered term frequency and the clustered inverse document frequency, the *clustered tf-idf* is given by

$$\text{ctfidf}_{j,d} = \text{ctf}_{j,d} \times \text{cidf}_{j,d}. \tag{B.3}$$

Qualitative description of an exemplary thread

To better illustrate our data, this section provides an in-depth qualitative description of one exemplary thread in the *SPON* discussion forum. Specifically, we consider a thread from the *Society* section that was originally attached to an article entitled *Why are people prone to believe in higher beings?*, published on January 1st, 2013. We first provide a translation of all comments in the thread, then we discuss structure and content in detail.

Table B2: Exemplary discussion thread

No.	Time	User ID	Comment
1	11:05am	User_1	1 + 1 = 2, 1 + 0 = 1, 0.75 + 0.25 = 1. How do you think the second equation should be interpreted? Who is Jesus, who is god? How the third one? How the countless others that are still imaginable? How would you face the existence of evil, knowing about the omniscience of God (The omniscent Creator forms the imperfect world? Why? So that it suffers? So that it can be screwed, what can hardly be denied (selling of indulgences, Luther and his wizards, Moses, Abraham, etc.)?) By the way, you wanted to have my opinion. Here you have what I think about the assumption of oneness.
2	12:23am	User_2	Both emanation models, the theistic and the scientific, are incomplete. If you leave out all the historic nonsense, then the only difference is that the theistic model presumes that the creation of World is based on a will.
3	12:38am	User_1	The existence of laws of nature, love, evil, belongs to Creation. Men as part of Creation are no puppets of God. They were provided with reason and conscience. But they often think that they are the actual Lords of Creation. They switch off their conscience and hold God responsible for the consequences. It's just as in the economy: privatize revenues, socialize losses.

Continued on next page

Table B2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
4	1:12pm	User_2	Ah, again the question of all questions. Well, reality is really very sad sometimes. But you could shoot a 24/7-soap opera: playing his (eternal) life: Sumerer. Plot: Eating the best food, then sex, then a bit of sleep, and then sending love comments with the computer (appeared in ep. 8 by flipping fingers) into the world. At the latest in ep. 4389 you want to step in and let Sumerer digress from the plot and ask what the shit is all about. Then the stage director smiles and says that he forgot to say that every actor has a free will, of course. In ep. 4390 then, Sumerer nibbles of the Tree of Knowledge, and later even more evil things come to his mind. But somehow I know this film – this story – already (at the very beginning of bible).
5	1:21pm	User_1	I guess this is why they threw Ashera, Jahwe's intimate partner, together with further heavenly legions, out of the temple and later palmed the sculpture of Virgin Mary off on him? How disappointed must God have been?
6	2:14pm	User_3	You are Gods. Evil things? Nonsense. Jealous gods drive each other to utter fury. Happens that one scratches eyes, breaks noses, bans or hijacks one's lover, blows up figures. Over the course of time, what happens is forgiven and forgotten. And merrily they proceed.
7	2:25pm	User_4	Only in a free economy do culture and civilization blossom (science), because the money is invested lest to lose value (like the flour in the jug in Thomas 97). There needs to be an anticipated liquidity payment if the money is not being invested, monthly, annual, or even daily. The fruit cannot generate further fruit by lending, because there is no interest any more. Yet there is no inflation, if the money is being invested in the medium or long run (in a bank, not in a jug of course), then there are no anticipated liquidity payments. It is such a system that releases the true productive, scientific, and social powers of men and bans sweet idleness. That requires of course, that one cannot sidestep to the monopoly of private property. These two monopolies lending og property and lending of money must be suppressed. In such an economy, Apple would be under control swiftly. Only this way, culture and civilization can develop sustainably.

Continued on next page

Table B2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
8	2:35pm	User_5	How is this droll utopia related to the topic? Your evangelical zeal for the prophet society would make you hero of every religious community, though.
9	3:48pm	User_6	Explain to me what "atheism" is supposed to be – I don't know. What substance, which meaning does this verbiage have?
10	16:07pm	User_6	Please compare the secular states of Europe, where every person can believe in whatever he or she wants, with the "theocracies". Then you realize where people are better off.
11	16:11pm	User_7	There is no "atheism". What's that supposed to be anyways?
12	16:33pm	User_8	If you are talking about the Hitler regime or Communism, you just offended all non-religois people. Non-religious people are neither (Neo-)Nazis nor Communists and therefore not responsible for the crimes of these regimes. Your problem is – in my opinion – that you do not understand the term "non-religious". For you there is only religious people. According to the principle: Everyone believes in something. That is not true. There are people who do not follow a leader. Neither a religious one, nor an Ideologist.
13	16:34pm	User_9	Then you gave whatever humanity does not know yet the name "God". Like the mathematician calls an unknown "x". But you didn't explain anything. Your x just has a different name now: God.

Continued on next page

Table B2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
14	17:06pm	User_10	One possible reason for "believe in higher beings" was not mentioned yet: the religious person does not only use gods to explain the world, he also wants to be protected. That's why gods have protective functions in many religions. Man prays to these gods so that they can help him. Sometimes it is ghosts, too: `http://www.spiegel.de/panorama/gesellschaft/aberglaube-in-thailand-wie-geister-das-leben-der-menschen-bestimmen-a-872769.html`. Modern Christians like to call believe in gods or spirits "superstition". And misses that the Holy Ones of Catholic Church are nothing else. Christianity knows evil spirits and demons, too. And the Vatican still offers classes for exorcism. Bottom line: modern Christianity has not developed far away from superstition of "primitive people". Even in modern Europe the world seems to be populated with invisible good and bad beings for religious people.
15	17:31pm	User_7	S.Freud was dealing with the topic and reaches two special points that he classifies as thought control through religion. He shows understanding for men's search for solace and comfort, and counts in religion, which is especially effective for granting the oldest and strongest wishes for protection and care via a mythologised father figure. Religion as illusion. Freud's main argument against religion is not, however, that it prohibits to enjoy life, but that it overdoes it and punishes resistance with oppression. Who submits to thought control is not able to reach the "psychological ideal, the primate of intelligence". Suppression of base instincts: Freud did not deal with scientific theology, in Roman-Catholic Austria or even Protestant theology, that resisted suppression of thought successfully since the end of the 18th century. He drew his knowledge about religion from his direct experience with Judaism to which he confessed. In his self-portrayal he writes in 1935: "Early absorption in biblical history, just after I mastered the art of reading, has, how I later realized, determined the direction of my thinking." His final work "The Man Moses and the monotheistic religion", published 1939, appreciates Judaism, because its strict rules and the suppression of basic instincts brought about the "triumph of intellectuality over sensuality".

Continued on next page

Table B2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
16	17:52pm	User_8	You are definitely mixing up cause and impact, because the "theocracies", especially in the Islamic world, did not cause the political and social crises, but were the consequence. Until a few decades ago, Iran was nearly as secular as the countries from the Arabic Spring were until recently. Whether it is the Mideast conflict, the lopsided support of Israel by the West, trade and oil, or simply the severe inferiority complex against the West that got the radical Islamists to power is a question that should be dealt with in other threads.
17	18:41pm	User_7	Now it's getting ridiculous. Just use Google if you still don't know it even after your umpteenth contribution to the forum, where you rattle off neo-atheistic points of view.
18	18:55pm	User_7	Slowly read again your quote above, you claim by yourself that religious theocracies are responsible if their states do not fare as well as we do in secular Europe. But whoever – just like you! – blames non-secular countries for the political and social grievances must be consequent and blame the non-religious people there for the existing grievances, or not? And wherever – like in Communist states – atheism becomes doctrine (look up Karl Marx!), then it does not help the atheists to hit and run and have nothing to do with the massacres that were committed in the name of humanism. Because these were nothing else than atheistic theocracies! Horrible crimes were committed in Christianity that Christians are being reminded of all the time. If I as a Christian am held reliable, you as atheist should be too. Any further questions?
19		User_11	One should positively mention the Egyptian Pharaoh Hatschepsut, who introduced the multi-day Opet festival. The beer, oh the beer flew like water. One reckons that the festival had a positive impact on fertility at the Nile, while, what is sad but true, Jahwe's bride was hijacked later on in Israel and he was lonely every since.
20		User_7	I don't know right know, but I think that some psychoanalyst once called the exile from Egypt "birth". Was that Freud or Jung?

Continued on next page

Table B2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
21		User_12	Adam and Eve could be cast out of paradise. That means that paradise has boundaries and is not endless. This means, that there is only a certain number of squared meters of paradise available. Who evangelizes is responsible for congestion. If you turn everyone to faithful, it's gonna be like subway in Tokyo up there. – free quote after Marc Uwe Kling, "The kangaroo manifest".

The discussion thread features 21 comments from 12 unique users (we replaced the original user names wit User IDs). All comments were written within a few hours on the date of publication of the underlying article. The comments vary in length: while some comprise just two or three sentences (e.g., comments 2 and 5), others are considerably longer (e.g., comments 7 and 15). All comments are somehow related to religion, which is the primary topic of the underlying article, but starting from comment 7, the discussion digresses towards political and economic issues, too. Some users reply to each other, but this is not always the case. E.g., comments 1 to 3 are seemingly unrelated to each other, but comment 4 is a direct reply to comment 3. Similarly, comments 6 and 7 are related to the general discussion in the thread but do not respond to any previous comments; comment 8, in contrast, is a reaction to comment 7. While many comments contribute to an overall (developing) discussion within the thread, some comments are just random (e.g., comment 18). We also observe that direct interactions between users are relatively short-lived: e.g., User_1 and User_2 have a brief "conversation" in the beginning of the discussion – although they do not always immediately react to each other – and User_7 and User_8 have a brief conversation towards the end. These two conversations are not related to each other. In sum, the path dependency of the discussion within the thread is rather limited.

Appendix C

Appendix to Chapter 3

C.1 Data

Table C1: Share of entities by type of Orbis IDs

	mean
Regular IDs	0.949138
DE*-IDs	0.012705
Foreign IDs	0.002176
Branch-IDs	0.036139

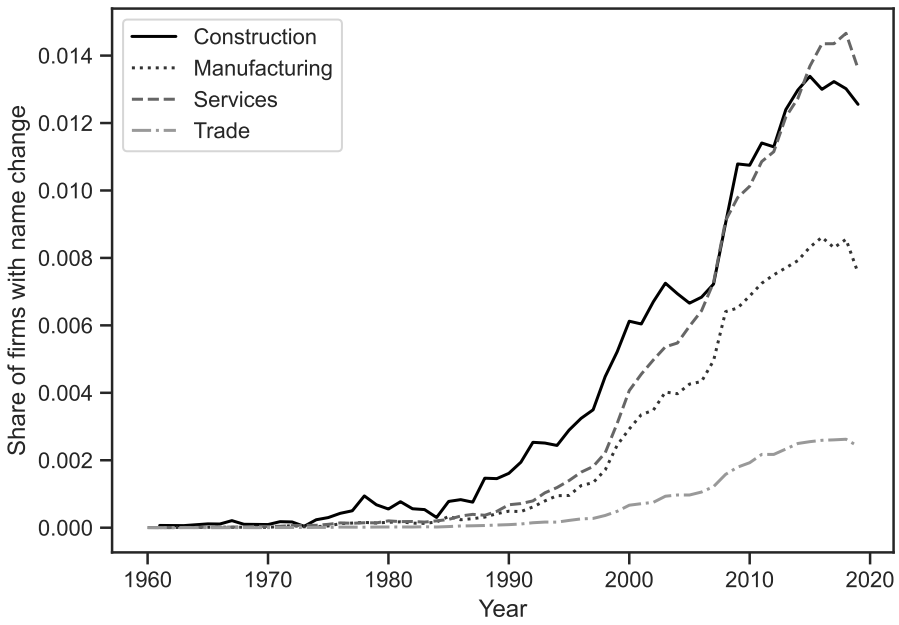
Note: *Regular IDs* take the form *DExxxxxxxx*. *DE*-IDs* indicate entities with some information that seemingly could not be matched to other variables. These usually contain no information that would be valuable for the research data set and are thus excluded. *Foreing IDs* refer to IDs that do not start with the country code *DE* but are nonetheless said to be located in Germany. *Branch-IDs* refer to IDs taking the form *DExxxxxxxxx-yyyy* where the last four digits are a running number starting with 1000 enumerating branches of the respective regular ID.

Table C2: Number of employees by survey

Survey	Employees
IBS-CON	65.28
IBS-IND	258.68
IBS-SERV	131.91
IBS-TRA	94.44
IVS-IND	442.25

Note: Contains the average number of employees by survey. Information about the number of employees is taken from the linked databases of the *BEP* and BIP and thus from the linked company databases. Thus, it is conditional on the entities being successfully linked.

Figure C1: Name changes in Orbis by business area



Note: Values are the number of name changes in a given year relative to the number of existing firms in that year. Includes only firms with available info on their date of incorporation. The increase in the share of name changes does not necessarily indicate a general trend but could be caused by Orbis being more likely to record name changes in the more recent years.

C.2 Linkage details

C.2.1 Preprocessing

Preprocessing steps:

- Transform information to right data type, i.e. *integer*, *string*, ... (includes transformation sector numbers to strings due to leading 0).
- Case folding (all lowercase)
- Replace German "Umlaute" (*äöü*) and other special letters respectively with *a*, *o*, *u*, and their ascii equivalents.
- Replace special characters
- Unify different spellings of "und" (and) such as *und*, *and*, *&*, and *+*.
- Special treatment for company names:
  - Remove and extract legal form via a set of regular expressions based on Schild (2016).
  - Identify special companies within a group such as *holding* or via regular expressions.

  - Extract a selection of other common terms such as *international*, *group*, *deutschland*, *Niederlassung*.
  - Create different versions of company name:
    1. Original string
    2. No whitespace and no special characters to better deal with compound words
    3. Array of tokens

- Telephone and fax:
  - Remove country code.
  - Parse into area code and number.
- Emails: keep only domain part.
- Location data:
  - Parse addresses into *street*, *number*, and *address supplement*.
  - Standardize different spellings of "straße" (street) and remove special characters.
  - Extract occurrences of zip codes from city.
  - Remove implausible zip codes.
  - Create 1-digit, 2-digit, 3-digit, and 4-digit sector identifiers
- Gather information into arrays:
  - Distinct alternatives (names, cities, addresses, phone numbers, and fax numbers)
  - Ranged address numbers (*"5-8"* $\rightarrow$ $\{5,6,7,8\}$)
- imputations/feature generation
  - Fill potentially missing primary info (e.g. phone number) with secondary info (alternatives).
  - Infer sector section from first two digits of sector identifier.
  - Infer manufacturing, retail/wholesale, construction, and services from sector section.
  - Use Deutsche Post Direkt (2019), to infer federal state, city, zip from other location information where uniquely possible
  - Double metaphone encoding.
  - Location word embedding.

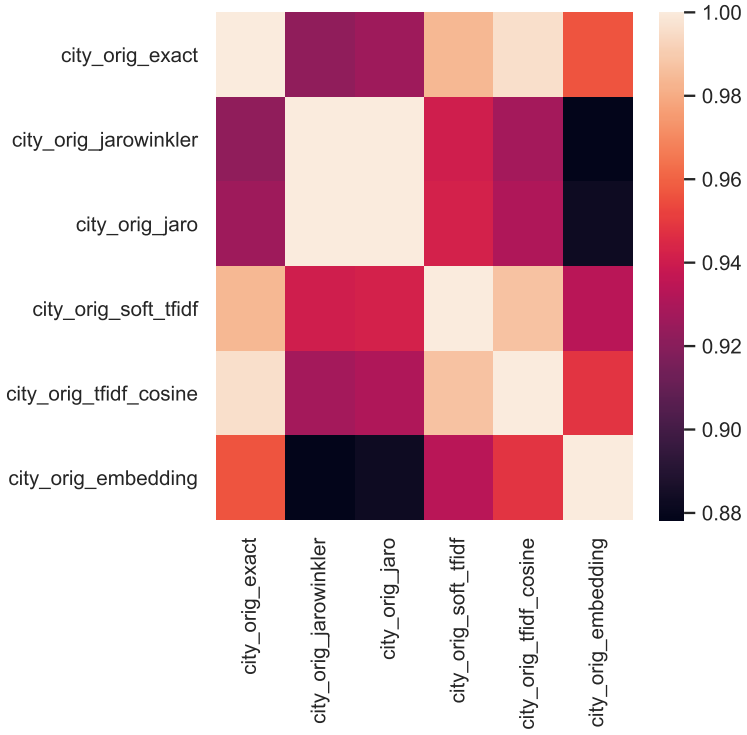
C.2.2 Comparison

Table C3: Most similar terms with FastText embedding vectors

Category	Term	Most similar
Not firm specific	Banane	Bananen, Ananas, banane
Not firm specific	Auto	Fahrzeug, Motorrad, PKW
Not firm specific	Firma	Tochterfirma, Herstellerfirma, Mutterfirma
Legal form	Aktiengesellschaft	Aktiengesellschaften, Kommanditaktiengesellschaft, Familien-Aktiengesellschaft
Legal form	AG	AG., AG-, AG7
Legal form	Gesellschaft	Gesellschaften, Gesell-schaft, Gesellschaft.
Legal form	GmbH	Co.KG, GbR, GmbHin
Legal form	e.v.	e.V., e.V, e.V.Im
Legal form	e.k.	e.k, ohg, e.K.
Location	Dresden	Chemnitz, Leipzig, Pirna
Location	Berlin	Potsdam, Berlin-Mitte, Charlottenburg
Location	München	Nürnberg, Augsburg, Starnberg
Location	Global	Gobal, global, European
Sector	Holding	Holdin, Holdings, Holding-Gesellschaft
Sector	Verwaltungsgesellschaft	Vermögensverwaltungsgesellschaft, Kapitalverwaltungsgesellschaft, Verwaltungsgesellschaften
Sector	Handelsgesellschaft	Außenhandelsgesellschaft, Handelsgesellschaften, Warenhandelsgesellschaft
Sector	Baugesellschaft	Wohnbaugesellschaft, Wohnungsbaugesellschaft, Union-Baugesellschaft
Sector	Bioscience	Biosciences, bioscience, Therapeutics
Sector	Schweisstechnik	Bewehrungstechnik, Schweißtechnik, Schweißtechnologie
Sector	Fertigteile	Fertigteil, Fertigteilen, Betonfertigteile
Sector	Invest	Investment, invest, Investments
Sector	Seniorenheim	Altenheim, Seniorenwohnheim, Pflegeheim
Name	Peter	Michael, Thomas, Andreas
Name	Meier	Müller, Baumann, Maier
Name	Schlenk	Schlenz, Schlenke, Schlenger
Colloquial name	Optimare	Maximare, Optimax, ComfortCtrl
Colloquial name	Airbus	Boeing, Airbusse, Airbus-Konzern

Note: Table shows which words are most similar to a selection of terms one can find in company names. Similarity is measured and terms are given for the pretrained FastText vectors used in this paper via the *most_similar* method from the python package *gensim*.

Figure C2: Correlation of metrics on the *city* field



Note: Cells show the Pearson correlation coefficient between different similarity metrics on the *city* field. The brighter the cell, the higher the correlation.

Table C4: Comparison features

Attribute	Transformation	Data Type	Method
Company name	Original	string	Exact
Company name	Name tokens	array	Monge-Elkan
Company name	Name tokens	array	Jaccard
Company name	Name tokens	array	Cosine (token)
Company name	Name tokens	array	SoftTFIDF
Company name	Name tokens	array	TFIDF-Cosine
Previous company name	Name tokens	array	Monge-Elkan
Previous company name	Name tokens	array	Jaccard
Previous company name	Name tokens	array	Cosine (token)
Previous company name	Name tokens	string	Smith-Waterman
Previous company name	Name tokens	string	Exact
Also known as company name	Name tokens	array	Monge-Elkan
Also known as company name	Name tokens	array	Jaccard
Also known as company name	Name tokens	array	Cosine (token)
Also known as company name	Name tokens	string	Smith-Waterman
Also known as company name	Name tokens	string	Exact
All company names	Name tokens	array	Multi Exact
All company names	Name tokens	array	Multi Jaro
Company name	Name without spaces	string	Exact
Company name	Name without spaces	string	LCS
Company name	Name without spaces	string	LCSSeq
Company name	Name without spaces	string	Smith-Waterman
Company name	Name without spaces	string	qgram
Company name	Name without spaces	string	cosine (ngrams)
All name variants	Array of all variants	array	Multi Exact
All name variants	Array of all variants	array	Multi Jaro
Company name segments (locations)	Name tokens	array	Multi Exact
Company name segments (locations)	Name tokens	array	Multi Jaro
Company name segments (first names)	Name tokens	array	Multi Exact
Company name segments (first names)	Name tokens	array	Multi Jaro
Company name segments (last names)	Name tokens	array	Multi Exact
Company name segments (last names)	Name tokens	array	Multi Jaro
Company name segments (sectors)	Name tokens	array	Multi Exact
Company name segments (sectors)	Name tokens	array	Multi Jaro
Company name segments (company name)	Name tokens	array	Multi Exact
Company name segments (company name)	Name tokens	array	Multi Jaro
City	Original	string	Exact
City	Original	string	Jaro-Winkler
City	Original	string	Jaro
City	Original	string	Soft-TFIDF
City	Original	string	TFIDF-Cosine
City	Original	string	Embedding cosine
City	Original	string	Frequencies
Postcode	Slices for each 1 digit to 5 digit	string	Exact
Postcode	Slices for each 1 digit to 5 digit	string	Frequencies
Street	Original	string	Exact
Street	Original	string	Jaro-Winkler
Street	Original	string	LCS
Street	Original	string	Smith-Waterman
Street	Original	string	Jaro
Street	Original	string	LSSSeq
Street	Original	string	Frequencies
Address number	Ranges	array	Multi Exact
Address number	Original	string	Levenshtein
Address number	Original	string	Frequencies
Address number	Zusatz	string	Exact
Address number	Zusatz	string	Levenshtein
Address number	Zusatz	string	Frequencies
Phone	Original	array	Multi Exact
Phone	Original	array	Multi Jaro
Email	Domain part	string	Exact
Email	Domain part	string	Jaro
Email	Domain part	string	Frequencies
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Exact
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Jaro
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Frequencies
WZ08	All sector classification slices for each 1 digit to 4 digit	array	Multi-Exact
Legal form	Categories	integer	Exact

Note: Table shows the similarity metrics that were used for each of the company attributes. The datatype *array* refers to a set of strings rather than one consecutive string. An example for this is {"Bayerische", "Motoren", "Werke"}. The methods *Multi Exact* and *Multi Jaro* refer to metrics where the respectively the maximum exact or Jaro score are measured in a cross comparison between all tokens of both records. The method *Frequencies* refers to a feature containing the relative frequency of the respective value of the records.

C.2.3 Classification

Table C5: Sizes of the training data

	Size	Share
Training data 1	8,307	0.56
Training data 2	3,561	0.24
Test data	2,968	0.20

Note: Training data 1 is used to train the individual components of the model, *training data 2* is used to then train the ensemble aggregator model, and *test data* is used to assess the quality of the model.

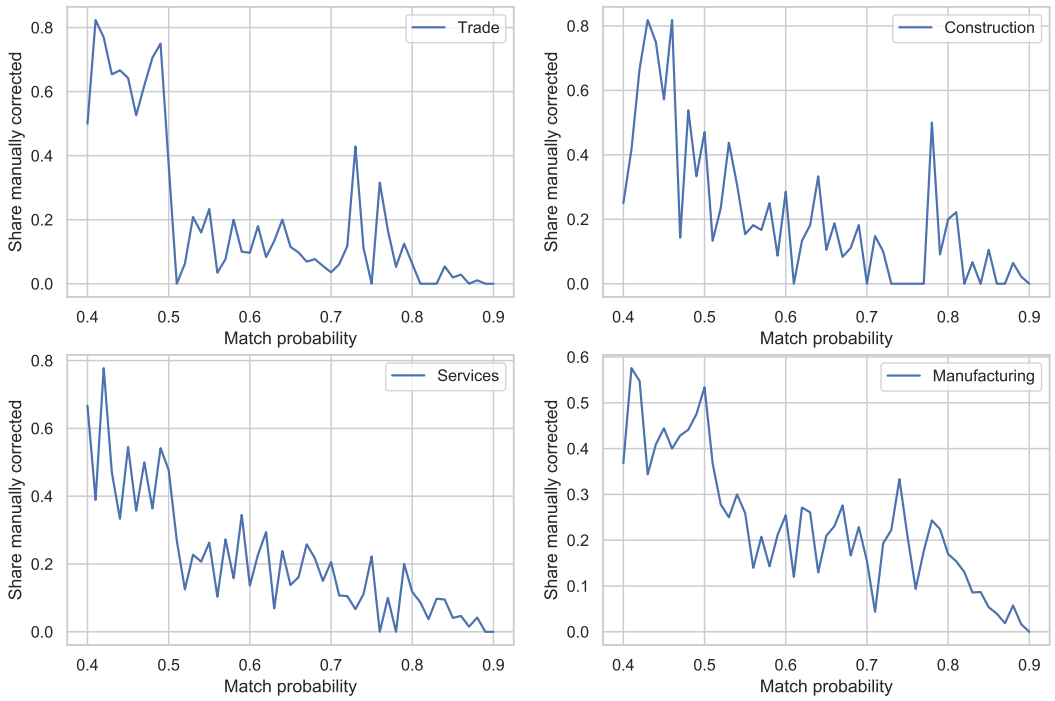
Table C6: Classification metrics

Estimator	Description	Accuracy	Precision	Recall	F1-Score
LogisticRegression	feature aggregation, continuous features	0.848	0.777	0.895	0.832
LinearSVC	feature aggregation, continuous features	0.850	0.774	0.907	0.835
MLPClassifier	feature aggregation, continuous features	0.865	0.828	0.857	0.842
XGBClassifier	feature aggregation, continuous features	0.877	0.823	0.901	0.860
LogisticRegression	feature aggregation, continuous features	0.850	0.779	0.896	0.834
LinearSVC	feature aggregation, continuous features	0.849	0.772	0.909	0.835
MLPClassifier	feature aggregation, continuous features	0.876	0.818	0.905	0.859
XGBClassifier	feature aggregation, continuous features	0.878	0.817	0.912	0.862
LogisticRegression	frequency weights, no missing data indicators	0.852	0.785	0.891	0.835
LinearSVC	frequency weights, no missing data indicators	0.849	0.776	0.899	0.833
MLPClassifier	frequency weights, no missing data indicators	0.877	0.817	0.909	0.861
XGBClassifier	frequency weights, no missing data indicators	0.886	0.847	0.888	0.867
LogisticRegression	continuous features	0.850	0.778	0.899	0.834
LinearSVC	continuous features	0.850	0.773	0.907	0.835
MLPClassifier	continuous features	0.865	0.809	0.887	0.847
XGBClassifier	continuous features	0.872	0.819	0.891	0.853
RandomForestClassifier	categorical features, binned continuous features	0.865	0.787	0.927	0.851
CatBoostClassifier	categorical features, binned continuous features	0.890	0.848	0.898	0.872
RandomForestClassifier	no missing data indicators, binned continuous ...	0.871	0.800	0.924	0.857
CatBoostClassifier	no missing data indicators, binned continuous ...	0.866	0.824	0.864	0.843
MLPClassifier	frequency weights, categorical features	0.882	0.848	0.876	0.862
LogisticRegression	PCA	0.861	0.798	0.892	0.843
LinearSVC	PCA	0.860	0.790	0.906	0.844
MLPClassifier	PCA	0.890	0.848	0.898	0.872
Ensemble		0.898	0.852	0.916	0.883

Note: Accuracy is the share of correct predictions, *Precision* is the share of correct prediction among the predicted matches, *Recall* is the share of actual matches predicted as match, and *F1* is the harmonic mean of precision and recall. *MLPClassifier* is a Multilayer Perceptron, i.e., a Neural Network. The *Ensemble* is a logistic regression that takes the predictions from the models above as inputs to make the final match prediction.

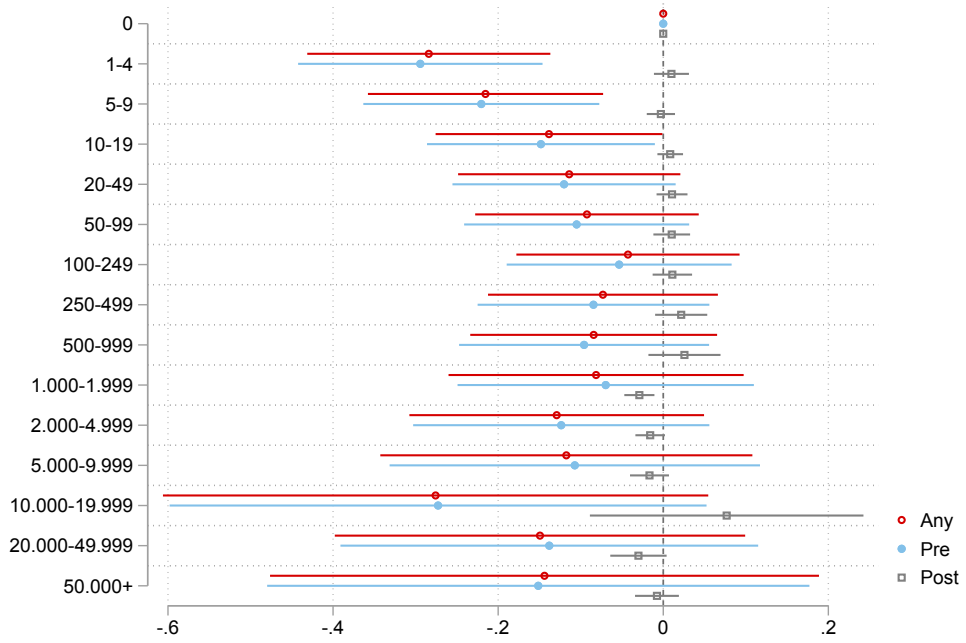
C.2.4 Postprocessing

Figure C3: Manual corrections by predicted match probability



Note: This figure shows the share of pairs that were manually corrected. Thus, for probabilities above 0.5, corrections reflect false positives corrected to negatives and for probabilities below 0.5, they reflect false negatives corrected to positives. Corrections were mostly conducted above the 0.5 threshold and the corrections below are based on a very small number of samples. The high number of corrections in the area below 0.5 is in part due to the fact that here, more or less obvious cases were selected, with a focus on false negatives, from a screening without further analysis of more complex cases.

Figure C4: Regression of match status on firm characteristics - size ranges



Note: Coefficients of a regression of match status on other characteristics from the ifo companies. Only coefficients on employment size range dummies are reported here, the others are shown in figure 3.8 for better visibility. Point estimates of a linear probability model are shown with 95% confidence intervals based on robust standard errors.

C.3 Access

Access to the data can be granted for purely scientific purposes at the *LMU-ifo Economics and Business Data Center* (EBDC) located at the ifo Institute in Munich, Germany. Because the data are confidential, access is only possible on-site on a protected workstation. The data contain no identifiers such as company name or address and it is prohibited to re-individualize individual companies. Only aggregated results, such as regression tables, can be exported and will be controlled by EBDC staff.

List of Figures

List of Tables

Bibliography

Abadi, Martin, A Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015) " TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," URL: `https://www.tensorflow.org/`.

Abowd, John M., Joelle Abramowitz, Margaret C. Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann M. Rodgers, Matthew D. Shapiro, and Nada Wasi (2019) *Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data*: US Census Bureau, Center for Economic Studies, URL: `https://ideas.repec.org/p/cen/wpaper/19-08.html`.

Abowd, John M and Lars Vilhuber (2005) "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business & Economic Statistics*, **23** (2), 133–152, URL: `https://doi.org/10.1198/073500104000000677`.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo (2022) "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics*, **40** (S1), S293–S340.

Agrawal, Ajay, Nicola Lacetera, and Elizabeth Lyons (2016) "Does standardized information in online markets disproportionately benefit job applicants from less developed countries?," *Journal of International Economics*, **103**, 1–12, URL: `http://dx.doi.org/10.1016/j.jinteco.2016.08.003`.

Akerlof, George A and Rachel E Kranton (2000) "Economics and Identity," *The Quarterly Journal of Economics*, **115** (3), 715–753.

Aldrich, Howard and Ellen R Auster (1986) "Even dwarfs started small: Liabilities of age and size and their strategic implications.," *Research in organizational behavior*.

Allcott, Hunt, Matthew Gentzkow, and Lena Song (2022) "Digital Addiction," *American Economic Review*, **112** (7), 2424–2463.

Altonji, Joseph G and Rebecca M Blank (1999) "Race and gender in the labor market," *Handbook of Labor Economics*, **3**, 3143–3259.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec (2013) "Steering User Behavior with Badges," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 95–106: ACM.

Anderson, Michael and Jeremy Magruder (2012) "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database," *The Economic Journal*, **122** (563), 957–989.

Antoni, Manfred, Katharina Koller, Marie-Christine Laible, and Florian Zimmermann (2018) "Orbis-ADIAB: From record linkage key to research dataset: Combining commercial company data with administrative employer-employee data," *IAB FDZ Methodenreport 04/2018 EN*, URL: `https://econpapers.repec.org/RePEc:iab:iabfme:201804_en`.

Archak, Nikolay (2010) "Money, glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on TopCoder.com," *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 21–30.

Arnold, David, Will Dobbie, and Peter Hull (2021) "Measuring Racial Discrimination in Algorithms," *AEA Papers and Proceedings*, **111**, 49–54.

Ash, Elliott, Daniel L Chen, and Arianna Ornaghi (2021a) "Gender attitudes in the judiciary: Evidence from US circuit courts," *Center for Law & Economics Working Paper Series*, **2019** (02).

Ash, Elliott, Ruben Durante, Maria Grebenschikova, and Carlo Schwarz (2021b) "Visual Representation and Stereotypes in News Media," *CESifo Working Papers* (9686).

Ash, Elliott and Stephen Hansen (2022) "Text Algorithms in Economics."

Athey, Susan (2019) "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence*: University of Chicago Press, 507–552.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019) "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, **109**, 65–70.

Athey, Susan and Glenn Ellison (2014) "Dynamics of open source movements," *Journal of Economics and Management Strategy*, **23** (2), 294–316.

Athey, Susan and Guido W Imbens (2019) "Machine learning methods that economists should know about," *Annual Review of Economics*, **11**, 685–725.

Bachmann, Rüdiger, Steffen Elstner, and Eric R Sims (2013) "Uncertainty and Economic Activity: Evidence from Business Survey Data," *American Economic Journal: Macroeconomics*, **5** (2), 217–249.

Backus, Matthew, Thomas Blake, Brad Larsen, and Steven Tadelis (2020) "Sequential Bargaining in the Field: Evidence from Millions of Online Bargaining Interactions*," *The Quarterly Journal of Economics*, **135** (3), 1319–1361.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey (2020) "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data," *Journal of Economic Literature*, **58** (4), 997–1044, URL: https://pubs.aeaweb.org/doi/10.1257/jel.20191526.

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong (2018b) "Social Connectedness: Measurement, Determinants, and Effects," *Journal of Economic Perspectives*, **32** (3), 259–280.

Bailey, Michael, Ruiqing Cao, Theresa Kuchler, and Johannes Stroebel (2018a) "The Economic Effects of Social Networks: Evidence from the Housing Market," *Journal of Political Economy*, **126** (6), 2224–2276.

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun (2020) "CEO Behavior and Firm Performance," *Journal of Political Economy*, **128** (4), 1325–1369.

Barach, Moshe A., Joseph M. Golden, and John J. Horton (2020) "Steering in Online Markets: The Role of Platform Incentives and Credibility," *Management Science*, **66** (9), 4047–4070, URL: http://pubsonline.informs.org/doi/10.1287/mnsc.2019.3412.

Bertrand, Marianne (2020) "Gender in the twenty-first century," *AEA Papers and Proceedings*, **110**, 1–24.

Bertrand, Marianne and Esther Duflo (2017) "Field experiments on discrimination," *Handbook of Economic Field Experiments*, **1**, 309–393.

Bilenko, Mikhail and Raymond J Mooney (2003) "Adaptive Duplicate Detection Using Learnable String Similarity Measures," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 39–48, New York, NY, USA: Association for Computing Machinery, URL: https://doi.org/10.1145/956750.956759.

Binette, Olivier and Rebecca C Steorts (2022) "(Almost) all of entity resolution," *Science Advances*, **8** (12), eabi8021, URL: https://www.science.org/doi/abs/10.1126/sciadv.abi8021.

Bitzer, Jürgen, Ingo Geishecker, and Philipp J.H. Schröder (2017) "Is there a wage premium for volunteer OSS engagement?–signalling, learning and noise," *Applied Economics*, **49** (14), 1379–1394.

Blackburn, Heidi (2017) "The status of women in STEM in higher education: A review of the literature 2007–2017," *Science & Technology Libraries*, **36** (3), 235–273.

Blau, Francine D and Lawrence M Kahn (2017) "The gender wage gap: Extent, trends, and explanations," *Journal of Economic Literature*, **55** (3), 789–865.

Bockstedt, Jesse, Cheryl Druehl, and Anant Mishra (2016) "Heterogeneous Submission Behavior and its Implications for Success in Innovation Contests with Public Submissions," *Production and Operations Management*, **25** (7), 1157–1176.

Bohnet, Iris (2016) *What works*: Harvard university press.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017) "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016) "Stereotypes," *The Quarterly Journal of Economics*, **131** (4), 1753–1794.

——— (2019) "Beliefs about gender," *American Economic Review*, **109** (3), 739–73.

Boudreau, Kevin J., Nicola Lacetera, and Karim R. Lakhani (2011) "Incentives and problem uncertainty in innovation contests: An empirical analysis," *Management Science*, **57** (5), 843–863.

Boudreau, Kevin J. and Karim R. Lakhani (2013) "Using the crowd as an innovation partner," *Harvard Business Review*, **91** (4).

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin (2022) "Social Media and Mental Health," *American Economic Review*, **112** (11), 3660–3693.

Brynjolfsson, Erik, Xiang Hui, and Meng Liu (2019) "Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform," *Management Science*, **65** (12), 5449–5460.

Burn, Ian, Patrick Button, Luis Munguia Corella, and David Neumark (2022) "Does Ageist Language in Job Ads Predict Age Discrimination in Hiring?," *Journal of Labor Economics*, **40** (3), 613–667.

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud (2020) "The Production of Information in an Online World," *The Review of Economic Studies*, **87** (5), 2126–2164.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017) "Semantics derived automatically from language corpora contain human-like biases," *Science*, **356** (6334), 183–186.

Caplan, Bryan (2018) *The case against education: Why the education system is a waste of time and money*: Princeton University Press.

Cavallo, Alberto and Roberto Rigobon (2016) "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, **30** (2), 151–178.

Charles, Kerwin Kofi and Jonathan Guryan (2011) "Studying discrimination: Fundamental challenges and recent progress," *Annual Review of Economics.*, **3** (1), 479–511.

Chevalier, Judith and Dina Mayzlin (2006) "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, **43** (3), 345–354.

Christen, Peter (2012) *Data Matching*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1–270, URL: `http://link.springer.com/10.1007/978-3-642-31164-2`.

Clark, Damon and Paco Martorell (2014) "The signaling value of a high school diploma," *Journal of Political Economy*, **122** (2), 282–318.

Claussen, JJrg, Pooyan Khashabi, Tobias Kretschmer, and Mareike Seifried (2018) "Knowledge Work in the Sharing Economy: What Drives Project Success in Online Labor Markets?," *SSRN Electronic Journal*.

Cohen, William W (2000) "Data Integration Using Similarity Joins and a Word-Based Information Representation Language," *ACM Trans. Inf. Syst.*, **18** (3), 288–321, URL: https://doi.org/10.1145/352595.352598.

Cohen, William W, Pradeep Ravikumar, and Stephen E Fienberg (2003) "A Comparison of String Metrics for Matching Names and Records," *Kdd workshop on data cleaning and object consolidation*, **3**, 73–78, URL: www.aaai.org.

Cohen, William W. and Jacob Richman (2002) "Learning to match and cluster large high-dimensional data sets for data integration," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 475–480.

Cuffe, John and Nathan Goldschlag (2018) "Squeezing More Out of Your Data: Business Record Linkage with Python," (18-46), URL: https://ideas.repec.org/p/cen/wpaper/18-46.html.

Currie, Janet, Henrik Kleven, and Esmée Zwiers (2020) "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, **110**, 42–48.

Dachwitz, Ingo (2016) "Analyse von Spiegel Online: So tickt Deutschlands größte Nachrichtenseite," *Netzpolitik.org*, URL: https://netzpolitik.org/2016/analyse-von-spiegel-online-so-tickt-deutschlands-groesste-nachrichtenseite/#netzpolitik-pw.

Davenport, Thomas H. and D.J. Patil (2012) "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*.

De Bruin, J (2019) "Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python," URL: https://doi.org/10.5281/zenodo.3559043.

Decarolis, Francesco and Gabriele Rovigatti (2021) "From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising," *American Economic Review*, **111** (10), 3299–3327.

Deming, David J. and Kadeem Noray (2020) "Earnings Dynamics, Changing Job Skills, and STEM Careers," *The Quarterly Journal of Economics*, 1965–2005.

Destatis (2008) "Klassifikation der Wirtschaftszweige, Ausgabe 2008," *German Federal Statistical Office*.

Deutsche Post Direkt (2019) "DATAFACTORY BASIC 2019 Q4."

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

Dissanayake, Indika, Nikhil Mehta, Prashant Palvia, Vasyl Taras, and Kwasi Amoako-Gyampah (2019) "Competition matters! Self-efficacy, effort, and performance in crowdsourcing teams," *Information and Management*, **56** (8), 103158, URL: `https://doi.org/10.1016/j.im.2019.04.001`.

Dissanayake, Indika, Jie Zhang, Mahmut Yasar, and Sridhar P. Nerur (2018) "Strategic effort allocation in online innovation tournaments," *Information and Management*, **55** (3), 396–406, URL: `https://doi.org/10.1016/j.im.2017.09.006`.

Doll, Hendrik, Eniko Gábor-Tóth, and Christopher-Johannes Schild (2021) "Linking Deutsche Bundesbank Company Data," *Technical Report 2021-05 – Version v2021-2-6. Deutsche Bundes-bank, Research Data and Service Centre.*

Dunn, Halbert L (1946) "Record linkage," *American Journal of Public Health and the Nations Health*, **36** (12), 1412–1416.

Easley, David and Arpita Ghosh (2013) "Incentives, gamification, and game theory: An economic approach to badge design," *ACM Transactions on Economics and Computation (TEAC)*, **4** (3), 16.

Eberle, Johanna and Michael Weinhardt (2020) "Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP)," *SSRN Electronic Journal*.

Ebraheem, Muhammad, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang (2017) "DeepER – Deep Entity Resolution," **11** (11), URL: `http://arxiv.org/abs/1710.00597%0Ahttp://dx.doi.org/10.14778/3236187.3236198`.

Ellemers, Naomi (2018) "Gender stereotypes," *Annual Review of Psychology*, **69**, 275–298.

Enders, Zeno, Franziska Hünnekes, and Gernot Müller (2022) "Firm Expectations and Economic Activity," *Journal of the European Economic Association*, **20** (6), 2396–2439.

Farronato, Chiara and Andrey Fradkin (2022) "The Welfare Effects of Peer Entry: The Case of Airbnb and the Accommodation Industry," *American Economic Review*, **112** (6), 1782–1817.

Fellegi, Ivan P and Alan B Sunter (1969) "A theory for record linkage," *Journal of the American Statistical Association*, **64** (328), 1183–1210.

Fisher, Robert J (1993) "Social desirability bias and the validity of indirect questioning," *Journal of Consumer Research*, **20** (2), 303–315.

Fiske, Susan T (2010) "Venus and Mars or down to Earth: Stereotypes and realities of gender differences," *Perspectives on Psychological Science*, **5** (6), 688–692.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz (2021) "The Effect of Social Media on Elections: Evidence from the United States," *NBER Working Paper Series* (No. 28849).

Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez (2015) *Countering online hate speech*: UNESCO Publishing.

Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg (2017) "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge," *Transactions of the Association for Computational Linguistics*, **5**, 529–542.

Garcia Martinez, Marian (2017) "Inspiring crowdsourcing communities to create novel solutions: Competition design and the mediating role of trust," *Technological Forecasting and Social Change*, **117**, 296–304, URL: `http://dx.doi.org/10.1016/j.techfore.2016.11.015`.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018) "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences of the United States of America*, **115** (16), E3635–E3644.

Garin, Andrew, Emilie Jackson, Dmitri K. Koustas, and Carl McPherson (2020) "Is New Platform Work Different from Other Freelancing?," *AEA Papers and Proceedings*, **110**, 157–161.

Gaulke, Amanda, Hugh Cassidy, and Sheryll Namingit (2019) "The effect of post-baccalaureate business certificates on job search: Results from a correspondence study," *Labour Economics*, **61** (September 2018), 101759, URL: `https://doi.org/10.1016/j.labeco.2019.101759`.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as data," *Journal of Economic Literature*, **57** (3), 535–574.

Giorcelli, Michela, Nicola Lacetera, and Astrid Marinoni (2022) "How does scientific progress affect cultural changes? A digital text analysis," *Journal of Economic Growth*, **27** (3), 415–452, URL: `https://link.springer.com/10.1007/s10887-022-09204-6`.

Glaeser, E. L., H. Kim, and M. Luca (2017) "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity," *NBER Working Paper Series* (No. 24010).

Glick, Peter and Susan T Fiske (2001) "An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality.," *American Psychologist*, **56** (2), 109.

——— (2018) "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism," in *Social cognition*: Routledge, 116–160.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (2021) "Social media, sentiment and public opinions: Evidence from #Brexit and #USElection," *European Economic Review*, **136**, 103772.

Gottapu, Ram Deepak, Cihan Dagli, and Bharami Ali (2016) "Entity Resolution Using Convolutional Neural Network," *Procedia Computer Science*, **95**, 153–158, URL: http://dx.doi.org/10.1016/j.procs.2016.09.306.

Gramlich, Tobias (2008) "Beschreibung der Verknüpfung der ifo-Konjunkturdaten mit der kommerziellen Firmendatenbank."

Grimmer, Justin and Brandon M Stewart (2013) "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, **21** (3), 267–297.

Grootendorst, Maarten (2022) "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*.

Gschwind, Thomas, Christoph Miksovic, Julian Minder, Katsiaryna Mirylenka, and Paolo Scotton (2019) "Fast Record Linkage for Company Entities," *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 623–630.

Gumpert, Anna, Henrike Steimer, and Manfred Antoni (2022) "Firm Organization with Multiple Establishments," *The Quarterly Journal of Economics*, **137** (2), 1091–1138.

Gusfield, D (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, EBL-Schweitzer: Cambridge University Press, URL: https://books.google.de/books?id=0fw5w1yuD8kC.

Hann, Il Horn, Jeffrey A. Roberts, and Sandra A. Slaughter (2013) "All are not equal: An examination of the economic returns to different forms of participation in open source software communities," *Information Systems Research*, **24** (3), 520–538.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman (2009) *The elements of statistical learning: data mining, inference, and prediction*, **2**: Springer.

Heckman, James J (1979) "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, 153–161.

Hernández, Mauricio A and Salvatore J Stolfo (1995) "The Merge/Purge Problem for Large Databases," *SIGMOD Rec.*, **24** (2), 127–138, URL: https://doi.org/10.1145/568271.223807.

Herzog, T. N., F. J. Scheuren, and W. E. Winkler (2007) *Data Quality and Record Linkage Techniques*, **1**, New York: Springer.

Hettiarachchi, Gayan Prasad, Nadeeka Nilmini Hettiarachchi, Dhammika Suresh Hettiarachchi, and Azusa Ebisuya (2014) "Next generation data classification and linkage: Role of probabilistic models and artificial intelligence," *Proceedings of the 4th IEEE Global Humanitarian Technology Conference, GHTC 2014*, 569–576.

Hirschberg, Daniel S (1977) "Algorithms for the longest common subsequence problem," *Journal of the ACM (JACM)*, **24** (4), 664–675.

Hitsch, Günter J, Alı Hortaçsu, and Dan Ariely (2010) "Matching and Sorting in Online Dating," *American Economic Review*, **100** (1), 130–163.

Hochreiter, Sepp and Jürgen Schmidhuber (1997) "Long Short-Term Memory," *Neural Computation*, **9** (8), 1735–1780, URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hönig, Anja (2010) "Linkage of Ifo Survey and Balance-Sheet Data: The EBDC Business Expectations Panel & the EBDC Business Investment Panel," *Schmollers Jahrbuch*, **130** (4), 635–642.

Horton, John J. (2017) "The effects of algorithmic labor market recommendations: Evidence from a field experiment," *Journal of Labor Economics*, **35** (2), 345–385.

Horton, John Joseph and Moshe Barach (2020) "How Do Employers Use Compensation History?: Evidence From a Field Experiment," *Journal of Labor Economics*, 709277, URL: https://www.journals.uchicago.edu/doi/10.1086/709277.

Hotelling, H (1933) "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, **24**, 417–441.

Hsueh, Mark, Kumar Yogeeswaran, and Sanna Malinen (2015) ""Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?" *Human communication research*, **41** (4), 557–576.

Huang, Peng and Zhongju Zhang (2016) "Participation in open knowledge communities and job-hopping: Evidence from enterprise software," *MIS Quarterly: Management Information Systems*, **40** (3), 785–806.

Huber, Kilian (2018) "Disentangling the Effects of a Banking Crisis: Evidence from German Firms and Counties," *American Economic Review*, **108** (3), 868–898.

IBS-CON (2019) *Ifo Business Survey Construction 1/1991 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-con-2019b: LMU-ifo Economics & Business Data Center.

IBS-IND (2019) *Ifo Business Survey Industry 1/1980 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-ind-2019b: LMU-ifo Economics & Business Data Center.

IBS-SERV (2019) *Ifo Business Survey Service Sector 10/2004-12/2019*, Munich, DOI: 10.7805/ebdc-ibs-serv-2019b: LMU-ifo Economics & Business Data Center.

IBS-TRA (2019) *Ifo Business Survey Trade 1/1990 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-tra-2019b: LMU-ifo Economics & Business Data Center.

Isele, Robert and Christian Bizer (2013) "Active learning of expressive linkage rules using genetic programming," *Journal of Web Semantics*, **23**, 2–15.

IVS-IND (2019) *Ifo Investment Survey Industry 1964-2019*, Munich, DOI: 10.7805/ebdc-ivs-ind-2019: LMU-ifo Economics & Business Data Center.

Jaccard, Paul (1912) "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1," *New Phytologist*, **11** (2), 37–50.

Jaeger, David A and Marianne E Page (1996) "Degrees Matter : New Evidence on Sheepskin Effects in the Returns to Education Author (s): David A . Jaeger and Marianne E . Page Source : The Review of Economics and Statistics , Vol . 78 , No . 4 (Nov ., 1996), pp . 733-740 Published by : The MIT Pr," **78** (4), 733–740.

Jaro, Matthew A (1989) "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **84** (406), 414–420, URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478785.

Jensen, Robert and Emily Oster (2009) "The power of TV: Cable television and women's status in India," *The Quarterly Journal of Economics*, **124** (3), 1057–1094.

Kasai, Jungo, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa (2020) "Low-resource deep entity resolution with transfer and active learning," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5851–5861.

Kässi, Otto and Vili Lehdonvirta (2019) "Do Digital Skill Certificates Help New Workers Enter the Market? Evidence from an Online Labour Platform," *CESifo Working Papers*, **7810**, 1–31, URL: https://read.oecd-ilibrary.org/social-issues-migration-health/do-digital-skill-certificates-help-new-workers-enter-the-market_3388385e-en#page1.

Kearney, Melissa S and Phillip B Levine (2015) "Media influences on social outcomes: The impact of MTV's 16 and pregnant on teen childbearing," *American Economic Review*, **105** (12), 3597–3632.

Kelleher, John D and Brendan Tierney (2018) *Data Science*: MIT Press.

Kite, Mary E, Kay Deaux, and Elizabeth L Haines (2008) *Gender stereotypes.*: Praeger Publishers/Greenwood Publishing Group.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017) "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019) "The geometry of culture: Analyzing the meanings of class through word embeddings," *American Sociological Review*, **84** (5), 905–949.

Kroft, Kory and Devin G. Pope (2014) "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist," *Journal of Labor Economics*, **32** (2), 259–303.

La Ferrara, Eliana, Alberto Chong, and Suzanne Duryea (2012) "Soap operas and fertility: Evidence from Brazil," *American Economic Journal: Applied Economics*, **4** (4), 1–31.

Lambrecht, Anja and Catherine Tucker (2019) "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Management Science*, **65** (7), 2966–2981.

Lee, Samuel, Nina Moisa, and Marco Weiß (2003) "Open Source as a Signalling Device - An Economic Analysis," *Working Paper Series: Finance & Accounting*, **102**.

Lemus, Jorge and Guillermo Marshall (2021) "Dynamic Tournament Design: Evidence from Prediction Contests," *Journal of Political Economy (forthcoming)*, URL: `https://www.journals.uchicago.edu/doi/10.1086/711762`.

Leppämäki, Mikko and Mikko Mustonen (2009) "Skill signalling with product market externality," *Economic Journal*, **119** (539), 1130–1142.

Lerner, Josh and Jean Tirole (2001) "The open source movement: Key research questions," *European Economic Review*, **45** (4-6), 819–826.

——— (2005) "The economics of technology sharing: Open source and beyond," *Journal of Economic Perspectives*, **19** (2), 99–120.

Levy, Ro'ee (2021) "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, **111** (3), 831–870.

Link, Sebastian (2018) "Harmonization and Interpretation of the ifo Business Survey's Micro Data," *CESifo Working Paper Series* (December), URL: `https://ideas.repec.org/p/ces/ceswps/_7427.html`.

——— (2020) "Harmonization of the ifo Business Survey's Micro Data," *Jahrbücher für Nationalökonomie und Statistik*, **240** (4), 543–555, URL: `https://www.degruyter.com/document/doi/10.1515/jbnst-2019-0042/html`.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart (2023) "Information frictions among firms and households," *Journal of Monetary Economics*.

LinkedIn (2020) "About; https://about.linkedin.com/de-de; accessed 2020-08-01," URL: `https://about.linkedin.com/de-de`.

Liu, Meng, Erik Brynjolfsson, and Jason Dowlatabadi (2021) "Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis," *Management Science*, **67** (8), 4665–4685.

Loster, Michael, Manuel Hegner, Felix Naumann, and Ulf Leser (2018) "Dissecting company names using sequence labeling," *CEUR Workshop Proceedings*, **2191**, 227–238.

Loster, Michael, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Dirk Thomas (2017) "Improving Company Recognition from Unstructured Text by using Dictionaries.," in *EDBT*, 610–619.

Luca, Michael (2016) "User-generated content and social media," in *Handbook of Media Economics*, **1**: Elsevier, 563–592.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess (2019) "Augmenting Pre-Analysis Plans with Machine Learning," *AEA Papers and Proceedings*, **109**, 71–76.

Lundberg, Scott M and Su-In Lee (2017) "A Unified Approach to Interpreting Model Predictions," in I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett eds. *Advances in Neural Information Processing Systems 30*: Curran Associates, Inc. 4765–4774, URL: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

Marjanovic, Sara, Karolina Stańczak, and Isabelle Augenstein (2022) "Quantifying gender biases towards politicians on Reddit," *PloS one*, **17** (10), e0274317.

Mason, Lowell G (2018) "A Comparison of Record Linkage Techniques," (November), 2438–2447.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014) "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, **104** (8), 2421–2455.

McCrary, Justin (2008) "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, **142** (2), 698–714.

Meier, Tabea, Ryan Boyd, James Pennebaker, Matthias Mehl, Mike Martin, Markus Wolf, Andrea Horn, T Meier, R Boyd, J Mehl, M Martin, M Wolf, and M Horn (2019) ""LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015."

Meyer, Bruce D. and Nikolas Mittag (2019) "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net," *American Economic Journal: Applied Economics*, **11** (2), 176–204.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (2018) "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig (2013) "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Monge, Alvaro E and Charles P Elkan (1996) "The field matching problem: Algorithms and applications," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 267–270.

Moore, Jamie C., Peter W.F. Smith, and Gabriele B. Durrant (2018) "Correlates of record linkage and estimating risks of non-linkage biases in business data sets," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, **181** (4), 1211–1230.

Mudgal, Sidharth, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra (2018) "Deep Learning for Entity Matching," *Proceedings of the 2018 International Conference on Management of Data*, 19–34.

Mullainathan, Sendhil and Ziad Obermeyer (2017) "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, **107** (5), 476–480.

Mullainathan, Sendhil and Jann Spiess (2017) "Machine learning: An applied econometric approach," *Journal of Economic Perspectives*, **31** (2), 87–106.

Müller, Karsten and Carlo Schwarz (2020) "From hashtag to hate crime: Twitter and anti-minority sentiment," *Available at SSRN 3149103*.

Murtagh, Fionn and Pedro Contreras (2012) "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2** (1), 86–97.

Neuscheler, Tillmann (2023) "'Wir brauchen bessere Daten'," *Frankfurter Allgemeine Zeitung*.

Newcombe, H B, J M Kennedy, S J Axford, and A P James (1959) "Automatic Linkage of Vital Records," *Science*, **130** (3381), 954–959, URL: `https://www.science.org/doi/abs/10.1126/science.130.3381.954`.

Newcombe, Howard B (1988) *Handbook of record linkage: methods for health and statistical studies, administration, and business*: Oxford University Press, Inc.

Newcombe, Howard B and James M Kennedy (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Commun. ACM*, **5** (11), 563–566, URL: `https://doi.org/10.1145/368996.369026`.

Ong, Toan C., Michael V. Mannino, Lisa M. Schilling, and Michael G. Kahn (2014) "Improving record linkage performance in the presence of missing linkage data," *Journal of Biomedical Informatics*, **52**, 43–54, URL: `http://dx.doi.org/10.1016/j.jbi.2014.01.016`.

Orman, Wafa Hakim (2008) "Giving it away for free? The nature of job-market signaling by open-source software developers," *B.E. Journal of Economic Analysis and Policy*, **8** (1).

Osterloh, Margit and Sandra Rota (2007) "Open source software development—Just another case of collective invention?," *Research Policy*, **36** (2), 157–171, URL: `https://linkinghub.elsevier.com/retrieve/pii/S0048733306001983`.

Pallais, Amanda (2014) "Inefficient hiring in entry-level labor markets," *American Economic Review*, **104** (11), 3565–3599.

Papadakis, George, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas (2019) "A survey of blocking and filtering techniques for entity resolution," *arXiv preprint arXiv:1905.06167*.

Pearson, Karl (1901) "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2** (11), 559–572, URL: https://doi.org/10.1080/14786440109462720.

Pedregosa, F, G Varoquaux, A Gramfort, Michel V., B Thirion, O Grisel, M Blondel, Prettenhofer P., R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, 2825–2830.

Pennebaker, James, Ryan Boyd, Kayla Jordan, and Kate Blackburn (2015) "The Development and Psychometric Properties of LIWC2015."

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001) "Linguistic inquiry and word count: LIWC 2001."

Peruzzi, Michele, Georg Zachmann, and Reinhilde Veugelers (2014) "Remerge: Regression-based record linkage with an application to PATSTAT," *Bruegel Working Paper*.

Philips, Lawrence (2000) "The Double Metaphone Search Algorithm," *C/C++ Users Journal*, **18**, 38–43.

Podsakoff, NP (2003) "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, **885** (879), 10–1037.

Poetz, Marion K. and Martin Schreier (2012) "The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?," *Journal of Product Innovation Management*, **29** (2), 245–256.

Qian, Kun, Lucian Popa, and Prithviraj Sen (2017) "Active Learning for Large-Scale Entity Resolution," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 1379–1388, New York, NY, USA: Association for Computing Machinery, URL: https://doi.org/10.1145/3132847.3132949.

Řehůřek, Radim and Petr Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50, Valletta, Malta: ELRA.

Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010) "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.," in *LREC*.

Sarawagi, Sunita and Anuradha Bhamidipaty (2002) "Interactive Deduplication Using Active Learning," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, 269–278, New York, NY, USA: Association for Computing Machinery, URL: https://doi.org/10.1145/775047.775087.

Sauer, Stefan and Klaus Wohlrabe (2020) *ifo Handbuch der Konjunkturumfragen ifo Handbuch der Konjunkturumfragen*.

Schäffler, Johannes (2014) "ReLOC linkage : a new method for linking firm-level data with the establishment-level data of the IAB," *FDZ-Methodenreport*, **5**.

Schild, Christopher-J., Simone Schultz, and Franco Wieser (2017) "Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification," *Technical Report 2017-01, Deutsche Bundesbank Research Data and Service Centre*, URL: http://dl.acm.org/citation.cfm?doid=2951894.2951896.

Schild, Christopher-Johannes (2016) "Linking "Orbis" Company Data with Establishment Data from the German Federal Employment Agency," *German Record Linkage Center Working Paper No. 2016-02*.

Schlimmer, Jeffrey C. and Richard H. Granger (1986) "Incremental learning from noisy data," *Machine Learning*, **1** (3), 317–354.

Seiler, Christian and Christian Heumann (2013) "Microdata imputations and macrodata implications: Evidence from the Ifo Business Survey," *Economic Modelling*, **35**, 722–733.

Shapley, L S (1953) "17. A Value for n-Person Games," in Harold William Kuhn and Albert William Tucker eds. *Contributions to the Theory of Games (AM-28), Volume II*, Princeton: Princeton University Press, 307–318, URL: https://doi.org/10.1515/9781400881970-018.

Smith, T F and M S Waterman (1981) "Identification of common molecular subsequences," *Journal of Molecular Biology*, **147** (1), 195–197, URL: https://www.sciencedirect.com/science/article/pii/0022283681900875.

Spärck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, **28** (1), 11–21, URL: https://doi.org/10.1108/eb026526.

Stanley, Marcus, Lawrence F Katz, and Alan B Krueger (1998) "Developing skills: What we know about the impact of American employment and training programs on employment, earnings and educational outcomes," 51, URL: http://scholar.harvard.edu/lkatz/publications/developing-skills-what-we-know-about-impact-american-educational-and-training-pro.

Stanton, Christopher T. and Catherine Thomas (2016) "Landing the first job: The value of intermediaries in online hiring," *Review of Economic Studies*, **83** (2), 810–854.

Struß, Julia Maria, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner (2019) "Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language," in *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 354–365, Erlangen, Germany: German Society for Computational Linguistics & Language Technology.

Tejada, Sheila, Craig A Knoblock, and Steven Minton (2001) "Learning object identification rules for information integration," *Information Systems*, **26** (8), 607–633, URL: https://www.sciencedirect.com/science/article/pii/S0306437901000424.

Tetlock, Paul C (2007) "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, **62** (3), 1139–1168.

Ukkonen, Esko (1992) "Approximate string-matching with q-grams and maximal matches," *Theoretical Computer Science*, **92** (1), 191–211, URL: https://www.sciencedirect.com/science/article/pii/0304397592901434.

Vilhuber, Lars (2020) "Reproducibility and replicability in economics," *Harvard Data Science Review*, **2** (4), 1–39.

Wang, Zhongmin (2010) "Anonymity, social image, and the competition for volunteers: a case study of the online market for reviews," *The BE Journal of Economic Analysis & Policy*, **10** (1), 1–35.

Watanabe, Kohei (2021) "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages," *Communication Methods and Measures*, **15** (2), 81–102.

Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer (2018) "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language," in *Proceedings of the GermEval*, Vienna, Austria.

Wilson, D Randall (2011) "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage," in *The 2011 International Joint Conference on Neural Networks*, 9–14.

Winkler, William E (1995) "Matching and record linkage," *Business survey methods*, **1**, 355–384.

Wolf, Markus, Andrea Horn, Matthias Mehl, Severin Haug, James Pennebaker, and Hans Kordy (2008) "Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count," *Diagnostica*, **54**, 85–98.

Wooten, Joel O. (2022) "Leaps in innovation and the Bannister effect in contests," *Production and Operations Management*, **31** (6), 2646–2663, URL: https://onlinelibrary.wiley.com/doi/10.1111/poms.13707.

Wooten, Joel O. and Karl T. Ulrich (2017) "Idea Generation and the Role of Feedback: Evidence from Field Experiments with Innovation Tournaments," *Production and Operations Management*, **26** (1), 80–99.

World Economic Forum (2018) *The Future of Jobs Report*, **31**, 164–173.

Wu, Alice H (2018) "Gendered language on the economics job market rumors forum," *AEA Papers and Proceedings*, **108**, 175–79.

Wu, Alice H. (2020) "Gender Bias among Professionals: An Identity-Based Interpretation," *The Review of Economics and Statistics*, **102** (5), 867–880.

Xu, Lei, Tingting Nian, and Luis Cabral (2020) "What Makes Geeks Tick? A Study of Stack Overflow Careers," *Management Science*, **66** (2), 503–1004.

Zhang, Xiaoquan (Michael) and Feng Zhu (2011) "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review*, **101** (4), 1601–1615.