Three essays in applied microeconomics : of norms and networks

Jan Sonntag

Three Essays in Applied Microeconomics

Of Norms and Networks

Jan Sonntag

Thesis presented for the degree of Doctor of Philosophy

Institut d'Etudes Politiques de Paris
Ecole Doctorale de Sciences Po
Programme doctoral en économie
Département d'Economie
Doctorat en sciences économiques

Three Essays in Applied Microeconomics: Of Norms and Networks

Jan Sonntag

Thesis supervised by
Emeric Henry, Associate Professor at the Department of Economics, Sciences Po
Defended on December 12, 2019

Jury:

Charles Angelucci	Assistant Professor, Columbia Business School
Johannes Boehm	Assistant Professor, Sciences Po
Béatrice Boulu-Reshef	Professor, University of Orléans *(referee)*
Maria Guadalupe	Professor, INSEAD *(referee)*
Emeric Henry	Associate Professor, Sciences Po *(advisor)*

Contents

Note to the reader

The three chapters of this dissertation are self-contained research articles and can be read separately. They are preceded by an introduction which summarizes the research presented in this dissertation. While I am the sole author of the first chapter, the second chapter is co-authored with Emeric Henry and the third chapter is co-authored with Johannes Boehm, which explains the changes between the "I" and "we" pronouns. The second chapter has been published with minor modifications as:

Henry, Emeric and Jan Sonntag. 2019. "Measuring image concern." *Journal of Economic Behavior and Organization*, 160:19–39

Acknowledgments

Seeing the results of my years of research compiled in these pages, I must admit that I am filled with a certain pride and a feeling of accomplishment. It also makes me feel enormously grateful to all who have helped me along the way and without whom I would not be writing these lines.

First of all I would like to thank my thesis advisor and co-author Emeric Henry. I am lucky to have an advisor who encouraged me to pursue my curiosity, even if the ideas I came up with may at times have looked a little unorthodox. Emeric left me the freedom to learn and explore while at the same time always being available to challenge ideas, give me extremely valuable feedback, and help me forge nascent ideas into coherent research projects. His clarity of thought continues to set the bar and I hope that some of it rubbed off on me.

I am also deeply grateful to my secondary advisors Johannes Boehm and Julia Cagé. What started with a nice coffee chat with Johannes the day before I officially started my PhD turned into a great collaboration and a true friendship. His ability to stay focused on the big questions while coming up with innovative ideas for research continues to amaze me, and I will miss being able to drop by his office to discuss research, politics, life, and literature. Julia not only offered valuable advice, guidance, and crucial encouragement especially on my fledgling first chapter, but also set the example of how research can stimulate the debate outside the ivory tower. Staying true to rigorous research and simultaneously pushing for impact "in the real world" is not an easy task, and the way she manages to succeed in both is a source of inspiration.

I would also like to thank the remaining members of my thesis committee, Charles Angelucci, Béatrice Boulou-Reshef, and Maria Guadalupe, for taking the time to go through my articles in such depth. Their insightful questions and remarks are invaluable to improving the research presented here.

I am very grateful to Andrea Prat who agreed to mentor me at Columbia, provided invaluable feedback on my work and helped me take full advantage of the rich academic life that New York has to offer.

In addition to those already mentioned, many professors have helped my co-authors and me to improve the research underlying my thesis with their questions, feedback and advice. I would like to thank Yann Algan, Leonardo Bursztyn, Pierre Cahuc, Thomas Chaney, Pierre-André Chiappori, Quoc-Anh Do, Roberto Galbiati, Sergei Guriev, Jeanne Hagenbach, Nicolas Jacquemet, Benjamin Marx, Adrien Matray, Tamar Mitts, Suresh Naidu, Ezra Oberfield, Florian Oswald, Bernard Salanié, Steven Salop, Andrey Simonov, and Florian Szücs. I would also like to thank seminar participants at Columbia, Princenton, Paris 1, and Sciences Po, as well as at the ASFEE 2015, AFSE 2018, EARIE 2018, RES 2018, and SED 2018. For the second chapter I would like to thank our editor at the Journal of Economic Behavior and Organization, Daniela Puzello, as well as three anonymous referees for their helpful comments.

I would also like to express my gratitude to Philip Kreisel and Alex Urban of #ichbinhier

who not only agreed to share data on their counterspeech activities with me, but also patiently answered my many question about their approach.

Special thanks also to Daniel Bauer, Vinodkumar Prabhakaran, and Chandrachud Basavaraj for their feedback on my natural language processing methodology. Carolina Melches provided excellent research assistance on my first chapter and Maxim Frolov was of great help running the experiments presented in the second chapter.

No one would be doing research at the department if we did not have the administrative support of Alain Besoin, Pilar Calvo, Claudine Lamaze, and Sandrine Le Goff – thank you.

I would also like to thank my colleagues at BCG who have supported my decision to do a PhD, especially Amadeus Petzke and Just Schürmann. I thank BCG for financial support.

One of the very reasons I wanted to do a PhD was my love for teaching. A special thanks to the students I taught during my PhD - I am always delighted when I occasionally get an e-mail from a former student and get to hear how they have been.

The journey would not have been worth it and I would not have learned half as much without my fellow PhD candidates past and present, many of whom have become dear friends. Thanks first of all to my cohort(s): Sophie Cetre, Pierre Cotterlaz, Florin Cucu, Etienne Fize, Arthur Guillouzouic Le Corff, Charles Louis-Sidois, Aseem Patel, Joanne Tan, Julien Pascal, Ludovic Panon, Camille Urvoy, and Jean-Baptiste Villain. Many thanks also to Edoardo Ciscato, Nicolò Dalvit, Pierre Deschamps, Anja Durovic, Jean-Louis Keene, Mario Luca, Elisa Mougin, Florence Nocca, Stefan Pauly, Luca Vernet, Max Viskanic, Riccardo Zago, Tyler Abbot, Vladimir Avetian, Victor Augias, Daniel Barreto, Oliver Cassagneau-Francis, Pauline Corblet, Edgard Dewitte, Marcos Diaz, Jean-Benoît Eymeoud, José Rodrigo Lopez Kolkovsky, Alaïs Martin-Baillon, Clément Mazet, Julia Mink, and Zidney Wong. I think the community we have built in the "combles" is a great one and certainly makes research more collaborative and fun. A special shout-out also to all PhDs I met (again) at Columbia, especially Seung-Ho Lee, Divya Singh, Qing Zhang, Louise Guillouët, Bhargav Gopal, Vinayak Iyer, and Haaris Mateen.

I am deeply grateful to have great friends and family I could always rely on to be there for me, cheer me up when needed and remind me that there is life outside academia. Thanks especially to Arno, Caro, Clemens, Edi, Felix, Flo, Franzi, Freddy, Hannes, Lennart, Niklas, Sarah, and Shari.

At the risk of sounding cliché, I do want to thank my parents, Edith and Gert. Their unwavering support, good questions, and advice have helped me overcome more than one hurdle during my thesis.

The best outcome of my PhD-years cannot me found on the pages of my thesis: the serendipitous discovery of meeting Catherine. There is no words to express what I owe to her and how grateful I am for having her in my life.

Introduction

This dissertation combines the fruits of three research projects revolving around two wider topics: social norms and production networks. Both these topics pertain to settings which give rise to externalities, turning them into areas of economic research that are relevant for the design of policies and incentives. Because the two topics are otherwise quite different from each other, I will describe each of them in turn.

The aim of the first two chapters is to further our understanding of how social norms shape our behavior. There is now a large literature documenting that social norms play an important role in explaining human behavior in a wide range of situations. Research that I will review in more detail in the main body of the thesis has identified social norms as a key driver of behaviors such as labor force participation, teamwork and effort provision, political participation, educational attainment, charitable giving and other forms of prosocial acts, to name a few. Social norms are already being leveraged in a number of settings to induce a desired behavior and help to overcome collective action problems, from voting to electricity consumption (Allcott (2011) and Gerber and Rogers (2009)). Indeed, norms-based interventions are attractive in many cases because they are often cheaper and easier to implement than monetary incentives.

Failing to understand how norms form, how they are sustained, and how they shape individuals' behavior can lead to costly inefficiencies through ill-designed policies or institutions. Far from being an afterthought, social norms can "crowd-out" monetary incentives, causing well-intentioned incentive designs to back-fire. Moreover, individuals have been shown to engage in punishment in order to sustain these norms, often at cost to themselves. Growing the body of research on social norms thus helps to anticipate these types of behaviors and take them into account in the design of contracts, institutions, and policies.

In modern economic models, social norms are usually conceptualized by maintaining the assumption of rational utility maximizing agents but modifying their utility functions. For example, agents in Bénabou and Tirole (2006) derive utility from behavior aligned with social norms through two channels: First, there is an agent's intrinsic motivation to take a certain action, a direct component in the utility function which yields a positive payoff associated with the action. Second, there is social image, a taste for the positive inference an outsider observing the agent's action would make about his or her character. Image concerns may indeed induce individuals to behave much more prosocially than they would if their behavior would not give them the benefit of observers' admiration or the weight of their disapproval.

This framework of thinking about social norms motivates my first two chapters. The first one investigates a specific modern day case study where social norms are leveraged in the fight against online hate speech to shed light on how norms shape political behavior more broadly. My contribution is to show that speaking out against hateful views is an effective way of deterring

further hate speech. The mechanism that most likely explains this effect is that vociferous contradiction in fact serves as a form of non-monetary punishment that communicates the presence or raises the salience of a social norm.

The second chapter focuses on the crucial role of image concerns in explaining the effect of social norms on behavior. Indeed, while there are now plenty of studies showing that image concerns affect people *on average*, we still know very little about which individuals specifically drive that effect. Taking into account the heterogeneity in image concerns, however, is important for leveraging social image and understanding its distributional consequences. Our lack of knowledge in this area is to a large extent due to the fact that there has been no standard way of measuring individuals' responsiveness to social image. Chapter two introduces a novel laboratory experiment designed to fill this gap. It generates an individual-specific measure of image concern, shows that there is substantial heterogeneity even in a small laboratory sample, and investigates how it correlates with other social preferences.

The final chapter of my thesis moves on from social norms and networks to production networks. In the modern economy, the idea of division of labor has been extended from workers to firms such that today's supply chains form complex networks spanning the globe. Understanding the characteristics and determinants of these networks is important both because they affect the extent to which local shocks propagate into aggregate fluctuations and because they impact individual firms' performance.

One key parameter of these networks is the boundaries of firms: supply chains are characterized by varying degrees of vertical integration. While specialization may offer efficiency gains, so can exploiting synergies between vertically related firms, the elimination of double marginalization or the alleviation of hold-up problems. However, vertical integration may also give rise to anticompetitive behavior or indeed be a motive for it. In the last chapter, I discuss one such mechanism, called vertical foreclosure, by which vertically integrating firms disrupt the supply of critical inputs to competitors. I leverage novel production network data to identify mergers and acquisitions between vertically related firms and asses to what extent these mergers affect the supply chains of their rivals.

Despite their apparent disparity, the three chapters share a few important characteristics. All three are motivated by an extant theoretical literature but are primarily empirical in nature. Underlying all three chapters are separate data collection efforts with the objective of bringing new data to longstanding problems. Indeed, each chapter contains elements that I hope can be used as building blocks for further research: chapter one demonstrates how deep learning techniques that are still barely used in economics can be leveraged in our discipline, chapter two describes an experiment to measure image concerns that can readily be included in other research designs, and the production network data at the heart of chapter three could prove to be useful also to investigate other questions than vertical foreclosure.

In the remainder of this introduction, I will provide a more detailed summary of each of the three chapters that can be skipped by readers intending to read the thesis in full.

Chapter 1 – Social norms and social media

The extent to which social media can shape people's views and ultimately actions has been the subject of intense debate over recent years as a rising tide of right wing populism in many Western societies coincides with the rise of social media. Of particular concern in this context has been its role in spreading false information – and hate speech. This is discomforting especially given mounting empirical evidence for the longstanding view that words ultimately lead to actions, establishing a causal relation leading from hate speech to hate crime.

In the first chapter of my dissertation, I ask whether norms-based interventions can help limit hate speech and therefore ultimately its negative consequences. One approach that could shape or communicate social norms governing the acceptability of certain statements is counterspeech, an intervention consisting of individuals themselves stepping in and contradicting hateful comments. Is this type of decentralized intervention effective at deterring hate speech and –more generally– to what extent can individuals influence each other's behavior in online debates?

To answer these questions, I use the Facebook pages of German language news media as a laboratory. There, a large grass-roots counterspeech group with more than 35,000 members intervenes each day on 1-2 media articles that received particularly large numbers of hateful comments. Members coordinate to write comments condemning hate speech on the selected articles and respond to hateful comments directly. These interventions do not go unnoticed: Due to the group's size and its focus on only a few interventions each day, counterspeech can take up a significant share of comments on targeted articles.

For six months, I collected all articles published on the Facebook pages of large German news media, as well as all user responses to them. I infer which of these comments contain xenophobic hate speech by training a deep neural network to classify the textual data, a methodological side contribution of my chapter as most textual analysis in economics so far has been restricted to much more basic techniques with important drawbacks. I combine these data with the counterspeech group's chat log which contains all articles that the group considered targeting with an intervention. This allows me to identify the treatment effects of interventions by comparing a treatment and a control group of individuals. The treatment group consists of individuals commenting on articles subject to an intervention. The control group contains individuals who were active on comparable "runner-up" articles which were considered by the group as targets for an intervention but ultimately were not chosen. I obtain these posts by restricting the sample of articles from the chat log to instances when the group faced a capacity constraint and had to choose between at least two ex ante similar posts in terms of their total number of comments, likes and hate comments. This ensures that treatment status of individuals is plausibly exogenous to their behavior.

The main result presented in this chapter is that the counterspeech interventions have a substantial but transitory moderating impact on individuals' future behavior. For about two weeks after an intervention, users in the treatment group are less likely to engage in hate speech

than the control group. The magnitude of this effect corresponds to a sizable 21% reduction. It is driven mainly by individuals who only occasionally spread hate speech and I find little effect on users who did so more than once a week. Moreover, targeted individuals tend to stay away from contentious debates prone to xenophobia. They reduce the number of comments they write or like, in particularly on articles pertaining to immigration-related topics.

Additionally, I show that counterspeech interventions change the composition of individuals that participate in online discussions. Targeted articles experience an increase in activity of up to 50% which is explained to an important extent by an influx of users who were not participants in the interventions. These individuals are less likely to have a previous history of hate speech and to make hateful comments. As a result, while the total number of hateful comments remains comparable to control posts, their share in the total activity created by individuals not participating in the intervention decreases by 3 percentage points.

Simple information provision is unlikely to account for my main findings. One could imagine that the effects stem from exposing individuals to new information which lead them to correct erroneous beliefs, for instance about crime rates among refugees, and adapt their behavior accordingly. However, only few counterspeech messages contain new pieces of information, suggesting that the interventions are not pure information treatments. In addition, the effects are temporary and smaller for individuals who get treated multiple times. These two facts seem difficult to reconcile with information provision as a driver of the changes in behavior.

Instead, I argue that my results are most consistent with interventions acting as non-monetary punishment communicating the presence of a social norm or increasing its salience. The mechanism through which social norms operate in this context does not seem to be that individuals infer a norm from the average behavior of others and align their own actions with it. For instance, I find no correlation between the share of counterspeech comments on an article and the effectiveness of the intervention. Rather, the behavioral change seems to be triggered directly by the disapproval expressed in the counterspeech messages – as if it was in fact a sanction for norm transgression. For example, I find that the effect of an intervention is strongest for those users who received a counterspeech message as a direct reply to their own comment, as opposed to a general comment to the article denouncing hateful comments. This response is consistent with the findings of the literature on non-monetary punishment, which highlights the role of sanctions in communicating social norms.

To the extent that one is concerned about hate speech reaching large audiences, inducing perpetrators to leave major news platforms is already an important step. However, one may be concerned that counterspeech interventions simply cause individuals to express hateful views elsewhere, potentially in environments prone to further radicalization. Of course, I only observe individuals when they comment publicly on news articles posted by German news media, but the data do contain news media that attract very high shares of xenophobic comments and no counterspeech interventions. I find no evidence that individuals shift their activity towards these outlets. This suggests that it is unlikely that there are large displacement effects to environments even further at the societal fringes.

My findings suggest that social norms on social media are a double-edged sword. While maintaining a dialog despite disagreements may help moderate toxic online debates, the mechanisms I describe may just as well work the other way around: unchecked hate speech triggers more hate speech. In this context, the fact that media companies seem to have an incentive to slant news towards articles that trigger hateful reactions in order to gain attention on social media is particularly worrying and highlights the special responsibility that editorial boards need to exercise.

Chapter 2 – Measuring image concerns

As I highlighted above, the extent to which individuals are responsive to social norms is determined in part by their image concerns. While individuals may have varying degrees of intrinsic motivation for pro-social behavior, even someone with no such motivation may follow a social norm if it confers sufficient honor on them – or not following the norm results in stigma or ostracism. There is a considerable empirical literature showing that image concerns matter for individual behavior *on average* and is reviewed in detail in the chapter. Yet, we know very little about the drivers and consequences of image concerns *at the individual level.* One of the main reasons for this gap in the literature is that we lack a systematic way of measuring the extent to which an individual is sensitive to the perception by strangers. Where studies do measure them at the individual level, they rely on highly context specific proxies making them unsuitable for generalization.

The second chapter of my dissertation is based on a joint paper with Emeric Henry in which we present a novel experimental game designed to measure image concerns at the individual level. It identifies image concern separately from other social preferences, such as altruism. The image concern game we propose involves three players: a dictator (he), a recipient and an observer (she). The dictator determines how much money to transfer to a lottery with two possible outcomes: success, in which case the recipient receives a given amount of money, or failure, in which case the recipient receives nothing. The more money the dictator transfers, the higher the chances of success. The dictator takes his decision knowing that the observer will be informed of the outcome of the lottery. Before the lottery is actually run, the dictator has to reveal his willingness to pay to remain anonymous (in an incentive compatible way), i.e. for his picture not to be revealed to the observer in case the lottery is a failure. The recipient never sees any pictures. The observer sees only the outcome of the lottery, not the amount the dictator actually transferred.

There are two main aspects that drive the structure of this game. First, image concern is easily measured by the willingness to pay to remain anonymous in case the recipient remains empty handed. Second, we show that if some reasonable properties of the utility function are satisfied, this measurement is independent of other social preferences including altruism. In case the dictator does not remain anonymous, the observer does not find out how much was contributed to the lottery, only that the lottery was a failure. Thus, the inference the observer

makes when she sees the picture is an updated belief on the characteristics of the dictator conditional on the fact that the lottery was a failure, and this belief cannot be conditioned on the actual amount transferred. Separating our measure from other social preferences is essential to understand the specific drivers of image concern and to show how it correlates with these other dimensions of preferences.

We find considerable heterogeneity in the strength of image concerns among participants in a laboratory experiment. About a third of our subjects are unwilling to pay any money in order to remain anonymous. While most participants are willing to pay *something*, another third of individuals pays quite high amounts. We show that few characteristics of the observer significantly impact the willingness to pay to remain anonymous. This is encouraging evidence of the portability of the setup. Nationality may be an exception: Non-French individuals pay significantly less for anonymity when facing other non-French observers and slightly more when observed by French observers. One possible interpretation is that non-French participants fear that due to prejudice, French observers will interpret a failed outcome of the lottery more adversely than non-French observers.

We then turn to the question whether image concerns are linked to other social preferences. After the image concern game, participants in our lab experiment played an infinitely repeated prisoner's dilemma game. In half of the experimental sessions the game was played with observers, in the others without. In sessions with observers, in addition to the two players of the prisoner's dilemma, we introduced a third player whose simple task was to observe the behavior of the other players and to rate their behavior after each round. This allowed us to document what actions are judged positively by the community and identify the prevalent social norm.

Using the repeated games, we first show that more image concerned individuals, when not observed, tend to cooperate less than others. We argue that this is evidence in favor of the fact that more image concerned individuals tend to be more selfish. Second, as mentioned above, comparing treatments run with observers to those without, we can show that more image concerned individuals correct their behavior in the direction of the social norm more than others – at least when they are observed by others.

This chapter of my dissertation introduces a systematic way to measure individuals' image concerns, validates the measure and starts exploring its drivers and consequences. It is designed to enable further research on the topic and I hope that other researchers will take up the experiment in their own endeavors. Due to the nature of the concept to be measured, the game is of course less portable than other games aimed at measuring social preferences, such as the trust game or the dictator game. However, the extensive analyses of observer effects and repeat measurements discussed in the chapter show that the setup can be simplified considerably in important aspects while preserving its key properties. For instance, in settings where retaliation outside the laboratory is unlikely, experimenters can use the recipient as an observer, rather than adding a third party observer. This can be useful, for instance in online or large sample settings. In the field, the process of taking pictures could be eliminated by asking the dictator to stand up if he loses anonymity. We attempted to test a maximum of threats to robustness

so that others don't have to and can build directly on our results.

Chapter 3 – Vertical integration and foreclosure

Vertical foreclosure arises when the producer of a bottleneck input integrates with one of its buyers and refuses to supply its downstream competitors. This can be the case when the integrating firms control access to infrastructure or technology that is required for producing the final good and uses this fact to limit other firms' ability to compete with it in the downstream segment. The competing buyers are said to be foreclosed.

There is a large theoretical literature examining the motivations for firms to engage in this behavior. The foundation for this modern, post-Chicago foreclosure theory is based on two seminal papers dating back almost thirty years ago: Hart and Tirole (1990) define the conditions under which firms can extend market power from the upstream to the downstream segment by engaging in foreclosure, while Ordover et al. (1990) stress foreclosure as a means to raise the input cost of rival downstream firms. On the empirical side, however, our understanding of whether firms use foreclosure as a business strategy in practice is limited to a small number of case studies in very specific settings. This can be explained by the fact that vertical relations are rarely observed but limits our ability to test the theories. Even less is known on how firms can respond to threats of foreclosure and mitigate their impact.

Perhaps as a consequence, the policy debate on the degree to which competition authorities should scrutinize vertical integrations is far from settled. In principle, many jurisdictions regard vertical foreclosure as violating competition law. In the United States, courts established a doctrine on foreclosure more than a century ago. Yet, how stringently authorities should enforce these rules continues to be debated. At the time of writing, both the US Department of Justice and the Federal Trade Commission are working on an update of their 1984 guidelines on non-horizontal mergers, and some legal and economics scholars call for "invigorating" competition policy (Baker et al. (2019)).

In line with the ongoing debate, enforcement of vertical merger cases is exceedingly rare in the US. In the last chapter, my coauthor Johannes Boehm and I ask if this is because of lax enforcement or due to vertical foreclosure barely occurring outside the minds of economists. What are the factors determining the prevalence of vertical foreclosure, and how severe are the consequences? How can firms threatened by foreclosure mitigate its impact?

To shed light on these questions, we examine the occurrence of vertical foreclosure across a range of industries and countries. We do this by exploiting a novel panel dataset on vertical relationships between large firms, both in the US and abroad. These data allow us to study whether buyer-seller relationships break following vertical mergers and acquisitions. We show that buyer-seller relationships are more likely to break when the supplier vertically integrates and the downstream merging firm is a competitor of the buyer – but not when the downstream merging firm is not a competitor of the buyer. Consistent with theories of vertical foreclosure, the former break probability is even higher when there is little competition in the upstream

industry. We rule out that these findings are explained by common industry-level (or industry-pair-level) shocks to merger activity or the break probability.

Of course, this correlation does not in itself imply that vertical foreclosure is taking place among the firms that we study. To give a causal interpretation to our findings, we address three threats to validity. First, we rule out reverse causality arising, for instance, from firms rescuing failing suppliers by acquiring them for supply assurance reasons. We show that our results still hold when we instrument suppliers' integration probability with capital outflows of mutual funds holding their equity – events that put downward pressure on these firms' stock prices for reasons unrelated to their performance.

Second, we argue that the correlation is unlikely to be driven by common shocks. We repeat our analysis on the sample of mergers that were rumored to occur but never materialized and find no impact. To the extent that these rumored integration events might be caused by the same unobserved shocks as actual integrations, they make for a good comparison group.

Finally, the links breaking might simply be the consequence of the integrating parties being so much more efficient that the buyer decides to exit the market. However, we show that firms whose *competitor* is vertically integrating and do not have a buyer-supplier relation with the integrating supplier do not experience a drop in sales, suggesting that strong synergies are unlikely to account for our results.

Two additional findings complete the last chapter. First, firms that have a foreclosure motive are more likely than others to integrate with a supplier. A foreclosure motive arises when a firm's supplier also sells to its competitor. Among active vertical relationships it is precisely between firms with such a motive that integration is most likely to occur. Second, we study the performance of firms in the wake of their supplier's integration. We document a sharp but transitory decline in sales among firms whose supplier integrates with one of their competitors. The sales drop is particularly pronounced for firms that do not already have a supplier from the same industry as the integrating supplier. This points to an important way for firms to mitigate the impact of foreclosure: diversification of the supplier base.

We interpret our results as supporting the view that vertical market foreclosure is occurring in the population of firms and relationships that we study. Of course, important caveats apply. First, because we do not have data on prices and quantities, we cannot make statements about the welfare effects of the integration events we study. Instead, we focus on the role of market structure and the impact on foreclosed firms. Second, the set of firms in our data consists mostly of firms that are either listed on exchanges or issue traded securities and is therefore a select sample. However, given that these firms and relationships will be more likely to be in the spotlight of antitrust authorities, we think that vertical foreclosure may also be prevalent outside the sample that we study.

References

Allcott, Hunt. 2011. "Social norms and energy conservation." *Journal of Public Economics* 95 (9-10): 1082–1095.

Baker, Jonathan B., Nancy L. Rose, Steven C. Salop, and Fiona M. Scott Morton. 2019. "Five Principles for Vertical Merger Enforcement Policy." *Georgetown Law Faculty Publications and Other Works,* no. 2148.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652–1678.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and so Should You." *Journal of Politics* 71 (01): 178.

Hart, Oliver, and Jean Tirole. 1990. "Vertical Integration and Market Foreclosure." *Brookings Papers on Economic Activity. Microeconomics* 1990 (1990): 205–286.

Ordover, Janusz A., Garth Saloner, and Steven C. Salop. 1990. "Equilibrium Vertical Foreclosure." *American Economic Review* 80 (1): 127–142.

Chapter 1

Social Norms and Xenophobia: Evidence from Facebook

1.1 Introduction

The ability of social media to shape people's views and actions has been the subject of intense debate in recent years, in particular in the aftermath of the US presidential elections and a sweeping rise of extreme right populism in many European countries. Its role as a platform for hate speech has received particular attention in this context. Recent reports of Facebook acting as a catalyst for violence against the Muslim minority in Myanmar put pressure on lawmakers and social media companies alike to rein in online hate speech (Reuters (2018)). Germany, for instance, adopted legislation that holds social media companies responsible for incendiary content posted by users on their websites. This paper evaluates the effectiveness of an alternative, decentralized approach to countering hate on social media: large scale counterspeech by users themselves.

A key motivation for curbing hate speech is that words will ultimately lead to actions. This concern is increasingly backed by empirical evidence. Yanagizawa-Drott (2014) and Adena et al. (2015) highlight the importance of mass media in fueling the genocides in Rwanda and Germany. Müller and Schwarz (2018a), 2018b argue that the prevalence of online hate speech increases the rate of hate crimes. Outside of the economics literature, there is evidence that hate speech has negative consequences for victims in terms of fear and participation in public debate (see Siegel (2018) for an overview).

If hate speech is so dangerous, how can it be kept in check? In principle, there are two broad approaches: centralized government intervention or a decentralized, more market based approach. The former has been adopted to some degree by many countries around the globe that have banned the most extreme forms of hate speech in order to protect minorities.[1] In modern democracies, however, the fundamental right of free speech limits the possible extent of such bans.

The second approach, favored for instance by social media companies, relies on counterspeech, a social control mechanism that requires users to step in, contradict, and speak out against hate speech in order to support victims and deter further transgressions. Organized counterspeech groups have formed and attracted sometimes large numbers of members that coordinate their efforts to counter hate speech. With the obvious risk of government intervention exceeding its aim and drifting into censorship, this decentralized approach seems of course very appealing.

The key question that this paper aims to answer is to what extent this bottom-up approach is effective and – more generally – to what extent individuals can influence each other's behavior in online debates. Does counterspeech cause targeted individuals to stop engaging in hate speech in the future? If so, what channels could explain this response? What happens to users who

[1] For example, German criminal law outlaws some of the most extreme forms of hate speech such as Holocaust denial, incitement to violence or civil unrest targeting protected groups. Similar laws exist in other European countries. In the US, on the other hand, the protection of First Amendment rights has generally outweighed concerns about the protection of minorities.

did not participate in the discussion before? Does counterspeech discourage new discriminatory messages?

In order to study these questions, I use the Facebook pages of German language news media as a laboratory.[2] In response to widespread hate speech on these pages, a large bottom-up counterspeech group was founded in late 2016 that attracted more than 35,000 members within a few months. The group intervenes each day on 1-2 media articles that receive particularly large numbers of hateful comments. An intervention consists of members coordinating to post comments condemning hate speech on the selected articles and to respond to hateful comments directly with the stated goal of countering and ultimately reducing the prevalence of hate speech. Due to the group's size and its focus on only a few articles, their interventions can take up a significant share of comments and were thus highly visible to anyone seeing the targeted articles on Facebook.[3]

To assess the impact of these interventions, I collected six months' worth of data on large German news medias' Facebook pages. The data include all posts by the media outlets, as well as all user comments, "likes", and replies to comments that were made in public and visible to anyone. I manually annotated several thousand of these comments and leveraged recent advances in deep learning for natural language processing to infer which of the millions of comments in my data contain hate speech. In addition, the counterspeech group shared a sanitized version of their leadership's chat logs with me which contain all articles that the group considered targeting with an intervention. This allows me to identify the treatment effects of interventions by comparing a treatment and a control group of individuals. The treatment group consists of individuals who were active on Facebook articles which were subject to an intervention. The control group contains individuals who were active on "runner-up" articles which were considered by the counterspeech group as targets for an intervention but ultimately were not chosen. I obtain these posts by restricting the sample of articles from the chat log to instances when the group faced a capacity constraint and had to choose between at least two ex ante similar posts in terms of their total number of comments, likes and hate comments. This ensures that treatment status of individuals is plausibly exogenous to their behavior.

The main result of this paper is that the counterspeech interventions have a substantial but transitory moderating impact on individuals' future behavior. For about two weeks after an intervention, users in the treatment group are 5.3 percentage points less likely to write or condone a xenophobic comment in a given week than the control group. After this period, individuals appear to revert to their initial behavior. Compared to the treatment group's baseline probability of engaging in hate speech during a given week of about 25%, this corresponds to a sizable 21% reduction. Individuals who only occasionally spread hate speech before the intervention alter their behavior the most, while I find little effect on users who did so more than once a week. The effect seems to be driven at least partly by targeted individuals staying

[2]Far from being solely a platform to connect with friends, Kennedy and Prat (2019) show that Facebook is now among the most influential news providers in many countries.

[3]For a survey of the group members' motivations to participate in these interventions, see Ziegele et al. (2019)

away from contentious debates prone to xenophobia. Following an intervention, they reduce the number of comments they write or like, in particular on articles pertaining to politics, immigration, and related topics. Their commentary activity on sports and the weather, on the other hand, increases slightly.

Additionally, I document that articles targeted by a counterspeech intervention see an influx of more moderate users participating in their discussion. Articles experience an increase of up to 50% in the number of comments and likes of comments, only parts of which is driven by participants in the counterspeech group: interventions trigger a ripple-on effect that attracts up to 20% more new users to the article who did not previously participate in any interventions. These additional users tend to be more moderate than individuals engaging on control posts and are less likely to make hateful comments. As a result, while the total number of hateful comments remains comparable to control posts, their share in the total activity created by individuals not participating in the intervention decreases by about 3 percentage points.

I discuss three possible channels which could explain my main results. First, interventions could simply provide additional information on the topic of the article. Individuals exposed to this kind of information treatment would then be able to correct erroneous beliefs, for instance about crime rates among refugees, and adapt their behavior accordingly. Second, individuals could infer social norms from the average behavior of other users discussing news articles. The intervention would then induce individuals with a taste for conformity with other users' behavior to not express xenophobic views. Finally, counterspeech could be a form of non-monetary punishment. The members of the intervening group write messages in which they publicly disapprove of the behavior of individuals who engage in hate speech. This could lead the latter to conform with the norm conveyed through these sanctions.

While the design of the natural experiment I am using for identification does not allow me to definitively answer which of these channels are at play, the findings seem to be most consistent with the interventions acting as non-monetary punishment. First, manually classifying a sample of counterspeech messages reveals that only a small fraction of them contain new factual information, suggesting that the interventions are not pure information treatments. In addition, the effects are temporary and smaller for individuals who get treated multiple times. This set of facts seems difficult to reconcile with information provision. Second, I find no correlation between the share of counterspeech in an article's discussion and the effectiveness of the intervention. This makes it unlikely that users infer a social norm from the distribution of others' comments on a given article and adapt their behavior to be more conform with others'. Third, I find that the effect of an intervention is strongest for those users who received a counterspeech message as a direct reply to their own comment, as opposed to a general comment to the article denouncing hateful comments. As most counterspeech messages are in fact expressions of disapproval of xenophobic views, this response is consistent with the messages acting as non-monetary punishment for norm transgression. The punishments' effectiveness could in principle be explained both by the messages inducing a behavioral response to shame, and by the messages communicating the presence of a social norm. While the results presented here

are consistent with both mechanisms, the social norms channel seems to be more relevant in light of previous findings in the literature on punishment.

Of course, I only observe the behavior of individuals when they comment publicly on news articles posted by German news media and I therefore cannot rule out that they simply express hateful views elsewhere, be it in private or outside Facebook. However, the data do contain articles from news media that attract very high shares of xenophobic comments and no counterspeech interventions and I find no evidence that individuals shift their activity towards these outlets. This suggests that it is unlikely that there are large displacement effects. Moreover, to the extent that one is concerned about hate speech being broadcasted to large audiences, inducing perpetrators to leave major news platforms may already be an important step.

Beyond the immediate setting I study, I argue that my results allow drawing more general conclusions. First, the fact that speaking up against hate speech has a sizable effect in a relatively impersonal online context suggests that the mechanism is likely not specific to this environment but holds true also elsewhere. This may indicate that contradicting hateful views is an effective intervention more broadly, both online and offline. Second, my findings shed light on how individuals respond to feedback by others more generally. They highlight that behavior is not hardwired by preferences, but that individuals' decisions can be affected significantly even by relatively isolated expressions of disapproval.

This paper contributes to the nascent literature on online hate speech, its drivers, and consequences. In a small scale experiment on Twitter, Munger (2017) uses false accounts to call out users for employing racist slurs and finds that they reduce their supply of hate speech temporarily. My results, on the other hand, rely on a large-scale natural experiment using actual interventions not administered by researchers which also allows me to track responses of users not directly targeted by an intervention. Müller and Schwarz (2018a) argue that anti-refugee comments on the Facebook page of Germany's populist right wing party cause an increase in the number of hate crimes in areas with high social media affinity. In the same vein, Müller and Schwarz (2018b) show that President Trump's racially charged Twitter comments may have led to a spike in hate crimes against Muslims. I build on their results and investigate the drivers of hate speech and a potential remedy.

More broadly, this research studies how social norms may affect behavior in an online context, thereby extending the existing literature on social norms in economics. Bénabou and Tirole (2006), 2012 develop a theoretical model in which individuals choose their actions in part as a response to what those actions tell others about them. The importance of individuals' concern for social norms in the context of political action has been documented by Bursztyn et al. (2017), who show experimentally that the election of Donald Trump made voters positively update their beliefs about the prevalence of xenophobic views in the population, causing xenophobes to reveal themselves more freely. Enikolopov et al. (2017) show that Russian protesters participated in demonstrations due to image concerns and provide a dynamic model of protest participation. In an experiment conducted in Pakistan, Bursztyn and Jensen (2016) show that men's decision to take a costly anti-American action depends in part on whether or not is

observable by a moderate majority of participants.

My work is also related to the literature on non-monetary punishment in the presence of social norms. Masclet et al. (2003) and Noussair and Tucker (2005) conduct a series of repeated public goods experiments in which they find that sanctions are effective at establishing and sustaining a norm of cooperation even when these sanctions do not directly affect participants' payoffs.[4] Xiao and Houser (2009) and Ellingsen and Johannesson (2008) show that in dictator games even the prospect of receiving written comments by recipients induces dictators to increase transfers. By varying the degree of publicity of sanctions, Xiao and Houser (2011) establish that punishments serve to express and raise the salience of social norms, rather than relying purely on a shaming effect. The results presented in this paper suggest that the findings on non-monetary punishment matter outside the laboratory as well.

On the methodological side, I extend the growing literature applying machine learning techniques to process text as data in economics (see Loughran and Mcdonald (2016) and Gentzkow et al. (2017) for an introduction and overview). In the context of political preferences, Groseclose and Milyo (2005), Gentzkow and Shapiro (2010), Jensen et al. (2013) and Gentzkow et al. (2016) use the US Congressional Record and apply simple, dictionary-based approaches to compute measures of polarization and slant based on the lexical differences between Republican and Democratic speeches. Relying more strongly on machine learning, Hansen et al. (2018) assess the impact of transparency on monetary policy deliberation. I build on more recent techniques in deep learning and use recurrent neural networks to identify hate speech in my corpus of Facebook comments.

The remainder of this paper is structured as follows: Section 1.2 introduces the data and provides descriptive statistics. Section 1.3 details the identification strategy. Section 1.4 reports the empirical evidence on the impact of counterspeech on the future behavior of treated users. The impact on aggregate user behavior on the targeted articles is contained in Section 1.5. Possible channels explaining these results are discussed in Section 1.6 before I conclude.

1.2 Data

1.2.1 Data sources

I combine data from three different sources. First, I collected roughly 12 million public user comments and 24 million "likes" on those comments pertaining to the 249,426 posts by 85 popular German language news media on Facebook between August 1st 2017 and February 1st 2018.[5] These data were publicly available through Facebook's API and include the comments' complete text, the time of writing, as well as a unique user ID that allows identifying users

[4] Gächter and Fehr (1999) similarly report an increase in cooperation when adding social feedback mechanisms to the public goods game, but only when participants were familiarized with one another before the experiment.

[5] This includes Facebook pages of news organizations with more than 60,000 followers, as well as a few manual additions of regional branches of the public broadcasting services and of the *Bild*, Germany's largest tabloid. Table 1.14 in the appendix contains the full list of media included.

across multiple Facebook pages.[6] In addition, I collected information on the exact time, the title and the "teaser" or description of the posts by news media, which are typically links to articles, pictures or videos.

Second, I identify all interventions of the counterspeech group *#ichbinhier* by manually collecting all calls to intervene on its internal Facebook group that is used to coordinate its roughly 35,000 members. In addition to the post ID required to identify on which post the group intervened, I also record the exact time of the call to intervene.

Finally, the organizers of the counterspeech group generously granted me access to a sanitized version of their internal chat that contains all mentions of urls along with a time stamp. As explained in more detail below, I use these urls to identify posts by news media on Facebook that the group considered as possible targets for their interventions.

1.2.2 Defining and identifying hateful comments

From the posts' and comments' raw text I extract the categorical variables for my econometric analysis. Most challenging in this respect is to find out which comments in the dataset contain hate speech and which ones do not.

Previous research has relied on dictionary-based techniques to identify hate speech. Munger (2017) searches the text for predefined racist slur-words, Müller and Schwarz (2018a) simply look for the word "refugee" on a right wing Facebook page to identify hateful messages. This approach proves to be insufficient in the present context for three main reasons. First, few users employ unambiguous racist slur words. Rather, they use mostly harmless language to convey racist ideas (e.g. "I think all Muslims should be forced to leave the country"). Second, the user base is not solely comprised of racists, so that occurrence of group names ("refugees", "Muslims", etc.) is not sufficiently predictive of hate speech but may in fact be used in a positive or neutral manner. Finally, users often misspell words or employ neologisms.

Instead, the approach chosen here relies on recent advances in machine learning. I first adopt a standard definition of hate speech to manually classify a set of comments and then train a deep learning algorithm on the manually categorized data. The algorithm then applies the learned categorization on the remaining data that were not manually classified.

The definition of hate speech applied here borrows heavily from Gagliardone et al. (2015) and as such is larger than the narrow definition of hate speech used by the German penal code, which views hate speech mainly as incitement to violence. A comment is counted as hate speech if it (i) insults, diminishes and/or approves discrimination of and/or violence against any group or group member based on religion, origin or ethnicity, (ii) generalizes (perceived) negative behavior or characteristics of individual group members to the group as a whole or (iii) questions intentions, honesty or ability of an individual based on group membership or spread clearly false information about the group with the intention to diminish, insult or spread

[6] No other user information was collected. Only comments that users made on public Facebook pages are included in the data.

prejudice.[7]

Using this definition, I manually classified a set of 15,000 comments with the help of a research assistant. About half of theses comments were sampled randomly from the entire dataset of comments, the remainder was selected by sampling from articles about refugees, immigration and crime – topics which attract a disproportionate amount of hate speech. This oversampling was necessary to insure a sufficient amount of xenophobic comments in the manually classified data as hate speech is, fortunately, a relatively rare phenomenon compared to the overall number of comments produced by users.

The hand-classified dataset thus obtained is used to train and evaluate various machine classifiers. The same reasons that rendered a keyword based approach infeasible in this context make classical bag-of-words machine learning approaches sometimes used in economics perform poorly as well (see Gentzkow et al. (2017) for an overview of these techniques). Instead, the most accurate classifier proves to be a Long Short-Term Memory algorithm similar to those often used in machine translation and text generation. These algorithms cope much better with the context dependency of the meaning of words, with misspellings and synonymy. A detailed description of the classifier is deferred to Appendix 1.A.

The classifier achieves a level of accuracy of 94.4 percent, which means that this percentage of comments is categorized correctly. Hate speech is relatively rare (about 3.4 percent in the entire dataset) making accuracy a somewhat uninformative metric. A classifier could simply never predict hate speech and still achieve 96.6% accuracy. More informative performance metrics are therefore the area under the receiver operating characteristic curve (93.4), as well as the classifier's sensitivity (56.9%) and specificity (97.3%).[8] While the classifier is certainly not perfect and slightly to conservative in predicting if any individual comment is hateful or not, it provides sufficient accuracy for the somewhat more aggregated analyses presented here.[9]

In addition to commenting on media posts, Facebook users can choose to react to other users' comments directly with an icon meant to convey different emotions, with Facebook's signature thumbs-up symbol ("like"), or with a free text comment (sub-level comment). I treat reactions to a hateful comments with a "like" or a heart-symbol as hate speech as the user clearly expresses agreement with the comment's content. After careful manual review of more than a thousand sub-level comments, I chose to exclude them from the textual analysis as they rarely contain unequivocal approval of the initial comment or new instances of hate speech.

Media articles are assigned to topic categories applying another machine classifier trained on a small manually annotated sample. While not strictly necessary for the analysis at the core of this paper, this intermediate step allows me to report more meaningful descriptive statistics.

[7]While I would have preferred conducting a more comprehensive analysis including homophobic and sexist hate speech, the data contain too few instances for robust statistical learning.

[8]The area under the receiver operating characteristic curve is the integral over the curve plotting the true-positive rate against the false-positive rate for each probability threshold of predicting positive outcome. Specificity is defined as the share of correctly predicted negatives in all negatives in the sample. Sensitivity is the share of correctly predicted positives in all positives in the sample.

[9]All key results presented in this paper are obtained using hate speech as the dependent variable so that classical measurement error should not introduce any biases.

1.2.3 Descriptive statistics

Users' responses to news articles

The media posts in the resulting dataset spawned reactions that differ substantially in their extent and nature depending on the posts' topic and the media outlet that produced it. The left panel of Table 1.1 shows the breakdown of posts (articles), the average number of user comments and likes, as well as the average share of these comments and likes containing hate speech by broad category of article topics. Most user reactions are attracted by articles that mention refugees. Despite the peak of the German "refugee crisis" of 2015 being long in the past during the observation period, there were almost daily articles about refugees on the social media pages of German media. These are also the articles that have the highest share of xenophobic comments. The second highest share of hateful comments is on miscellaneous articles that mostly cover crimes and lead many users to draw hasty conclusions about the origins of perpetrators.

These correlations are not merely driven by larger outlets producing more populist content. In Appendix 1.C.1 I report the results of regressions predicting activity levels and the share of hateful comments by topic-dummies and outlet fixed effects. It shows that the positive correlation between migration-related articles and the number of both hateful and other comments and likes is highly significant even when exploiting only variation within media outlets. Moreover, the share of xenophobic comments remains a strong, positive predictor of the number of comments and likes on an article, even when including a rich set of fixed-effects.[10]

The right panel of Table 1.1 shows how user activity is distributed by topic category. Users who comment on politics and on immigration related articles do so most actively, with an average of 10.9 and 7.9 comments and likes on these topics. They are also the most likely to have written or liked at least one xenophobic comment on an article on these topics. Overall, the average individual in the sample has written or liked a total of 10.1 comments and 9.6% of these users have written or condoned a xenophobic comment at least once.[11]

Both the total number of responses to articles and the share of xenophobic responses are highly volatile over time. Figure 1.1 suggests that some of this volatility can be attributed to specific events. For instance, the terror attacks in Barcelona on August 17, 2017 were followed by a large spike in the share of xenophobic comments, while the day of the German federal elections saw a spike in the overall number of comments and likes. Other peaks and troughs seem to be driven by multiple smaller coinciding events.

Individuals' activity

As is to be expected, the distribution of user activity is highly skewed. The vast majority of users only makes rare appearances on the public comment sections of major news media, on

[10] This suggests that outlets may have an incentive to produce news stories that will trigger xenophobic comments if they want to get more activity to their pages. I elaborate on this issue in the conclusion.

[11] Table 1.15 in the appendix provides the same breakdown by news media instead of topic.

Table 1.1: Posts, user activity and hateful comments by article topic

	Articles			Users		
	# Articles	Avg. activity	% Xen.	# Users	Avg. activity	% Xen.
Other	98,657	125.6	2.9	2,696,237	4.6	5.2
Refugees / Foreigners	21,071	270.2	14.8	718,966	7.9	24.1
Politics	51,955	208.3	4.3	995,894	10.9	13.4
Business	16,026	94.1	3.1	502,008	3.0	5.7
Miscellaneous	45,236	117.0	7.7	1,055,183	5.0	12.1
Sports	12,036	81.4	2.1	390,597	2.5	3.9
Weather	4,445	64.6	1.5	186,589	1.5	2.0
Total	249,426	148.2	3.2	3,645,980	10.1	9.6

Note: For each topic category, the columns in the first supercolumn report the total number of articles, the average number of comments and likes per article, the average share of xenophobic comments and likes respectively. The second supercolumn reports the total number of individual users who were active on the topic, users' average number of comments and likes and the share of users who wrote or liked at least one hateful comment. "Miscellaneous" contains articles about accidents, crimes and disasters. "Other" is a catch-all category containing articles about science, human interest stories, celebrity news, cooking, etc.

average 1.3 times per week (Figure 1.2). This is true despite the fact that I only observe users in my data that commented publicly at least once.

A concern often raised in the popular press is that social media debates are often waged by bots. Inspecting the high activity users, I find little evidence that would suggest that these profiles are robots: despite being very active, there speed is not superhuman, their messages are not simple copy-pastes or highly similar messages and often contain responses to other users that would be very difficult to automate. Their German contains typos, but is not excessively error ridden. Overall, I find few longer duplicate comments in my database that would suggest the presence of bots even among the less active users. This is in line with the fact that Facebook has more stringent identity verification processes in place than other social media companies making it relatively more costly (but not impossible) to create fake profiles that can be automated.[12]

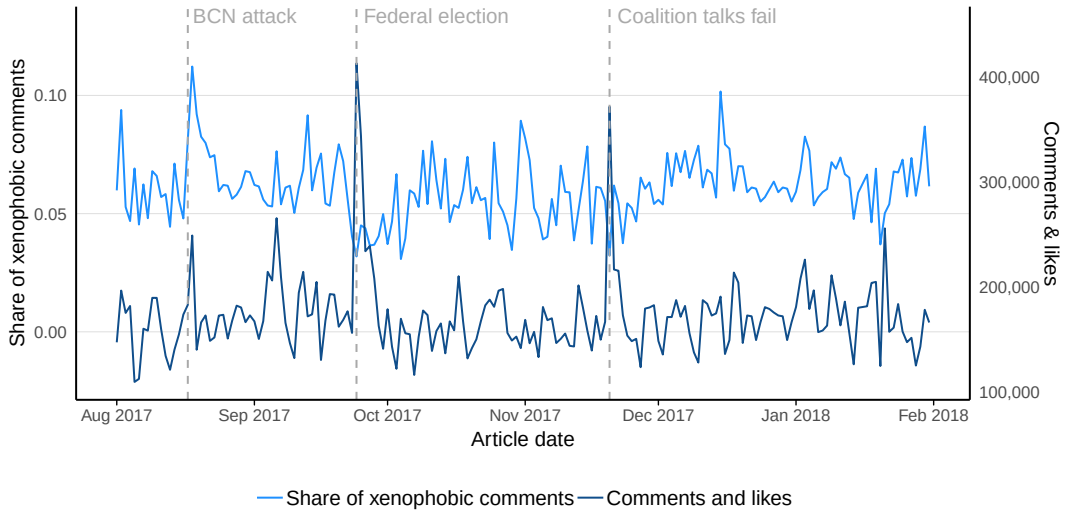
Counterspeech interventions

Figure 1.3 shows that the counterspeech group was quite successful at staging large interventions during the observation period. Although only relatively few posts were targeted, the majority of the 315 interventions managed to achieve a share of more than 20% of the overall comment section of that post. The average intervention involved 84 members of the counterspeech group and large ones more than a thousand. This lends credibility to the argument that these interventions did not go unnoticed by users who saw or commented on the news article on Facebook.

More generally, individuals who comment on news articles seem to also read and respond to the comments of others. One fifth of the users who write a comment on an intervention post

[12]I can only identify 154 profiles that are likely bots by looking for longer duplicate messages, users that post mainly links or that "never sleep". Excluding them does not change the analyses presented here.

Figure 1.1: User comments, likes and share of hate speech



Note: This figure reports the daily number of comments and likes (dark blue, right scale) and the xenophobic share therof (light blue, left scale) for each day in the observation period. For illustration, a few key dates are highlighted: the terrorist attack in Barcelona (Aug 17, 2017), the German federal elections (Sep 24, 2017) and the day the talks between major parties to form a new German government failed (Nov 20, 2017).

also like or comment on another user's comment. Among those whose comments received a reply by another user, at least 41% write another reply to that comment. Anecdotally, one can see users responding to the interventions in sub-level comments and engaging in (not always friendly) discussions.

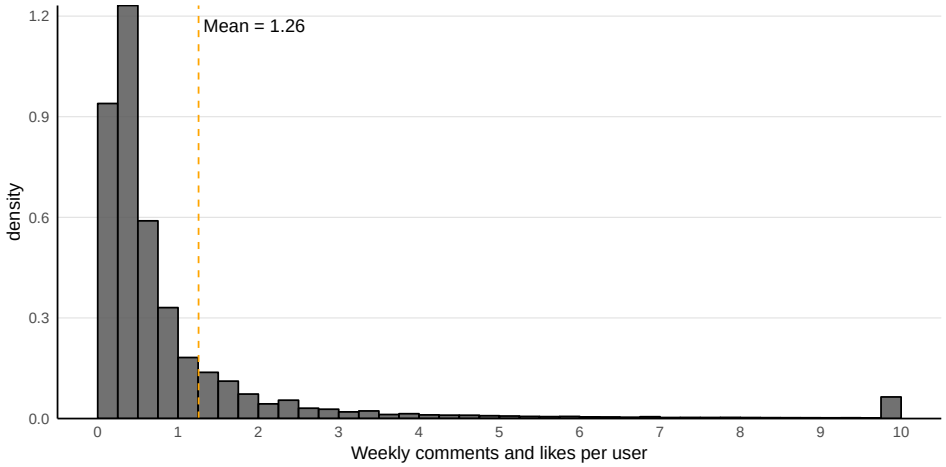
In order to provide a better sense of the content of the counterspeech messages, I manually classified approximately 600 comments into different content categories. The results of this exercise are reported in Figure 1.4 and in more detail in Appendix 1.E. I considered both top-level comments that were written as a comment directly on the article and sub-level comments which are responses to other users' comments.

The largest group of comments contains common sense arguments against xenophobic views. About a quarter of the top-level comment express plain disagreement with hateful views without providing additional information or arguments. This share is much lower for responses to other users' comments. Only 14% of top-level comments contain new pieces of information, such as statistics or a link to further reading on a topic. Social norms are invoked in about 11-13% of the comments and mostly in the form of injunctive norms. A minority of comments contain ad hominem attacks on other users such as expressions of doubt about their intellectual capacity or social standing. Outright insults or foul language are the exception. Equally rare are threats to report another user to Facebook or authorities. Among the remaining comments, many call for a more fact-based debate more generally or complain about the way the article is slanted by journalists to trigger hateful reactions by users.[13]

The types of articles and media outlets that were targeted by the counterspeech group are

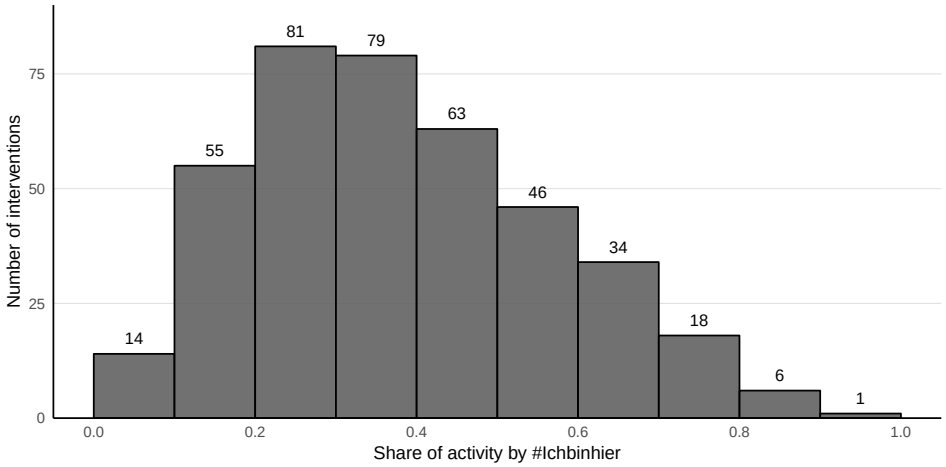
[13] Among the responses to other users' comments, there is also a large share which is difficult to interpret as they refer to previous sub-level comments in often lengthy discussions.

Figure 1.2: Histogram of the number of weekly comments and likes by users



Note: Histogram showing the distribution of the weekly number of comments an likes by user for all users active during the observation period whose first and last activity are at least five days apart.

Figure 1.3: Counterspeech interventions by share of comments in the post



Note: Histogram showing the distribution of the share of activity attributable to members of the counterspeech group *#ichbinhier* on posts the group decided to intervene on.

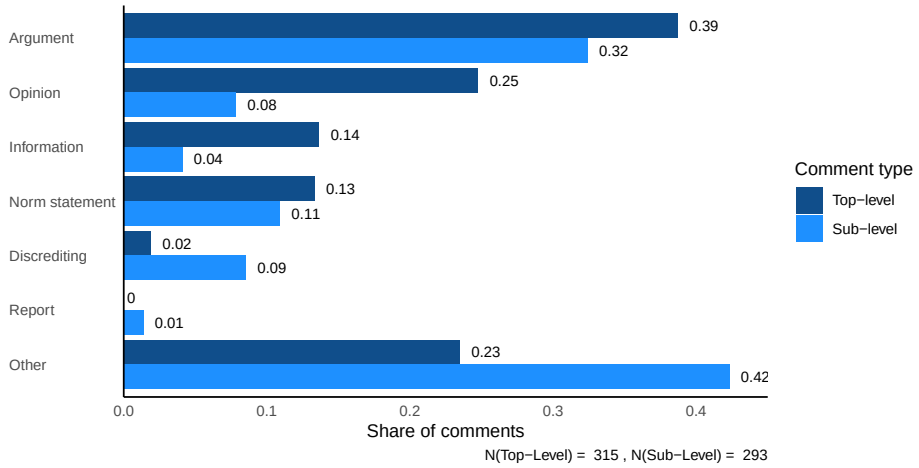
discussed in the next section after I have introduced the identification strategy.

1.3 Identification Strategy

The aim of this paper is to identify the causal effects of the counterspeech intervention both on the future behavior of users that were affected by them as well as the overall incidence of hateful comments on the targeted articles. Doing so requires constructing a credible counterfactual of what would have happened in absence of the intervention. To this end, I exploit the specific way in which these interventions were carried out.

The counterspeech group counted around 35,000 members at the time the data collection started but its organization was and still is highly centralized and revolves around a small

Figure 1.4: Content of counterspeech comments



Note: Manually classified comments by participants in counterspeech interventions by type of comment and content of the message. A comment can be assigned to multiple categories. *Argument* contains comments with common-sense based arguments against xenophobic statements, *Opinion* contains comments in which the author voices disagreement with hateful comments without providing arguments, *Information* counts comments that contain a new factual element to the debate, such as a statistic, *Norm statement* contains comments in which author invokes social norm, injunctive or descriptive, *Discrediting* contains ad hominem attacks, *Report* contains threats to report another user to site moderators, Facebook or the police. A more detailed breakdown and examples are available in appendix 1.E.

group of volunteers.[14] Every day, a handful of moderators monitors major German language news media on Facebook and picks an average of 2.1 posts a day on which to intervene. They specifically look for articles that were posted within the last two hours, attracted a lot of hateful comments and accumulated high numbers of comments and likes.[15] Once the moderators decide to intervene on a given post, they write a message on the board of the Facebook group so that it becomes visible to its thousands of members. Those who wish to follow the call to action can then write comments on the targeted article in a coordinated manner.[16] For concreteness, Appendix 1.D contains the timeline of one specific intervention as an illustrative example.

While the group's intervention is not random per se, I argue that it has two key features that allow for the identification of causal effects. First, the group faces a capacity constraint: it only intervenes on one post at a time and only on a few posts per day in order not to divide its resources among too many interventions. Especially at times when news break which spawn many hateful reactions, this means that the group is forced to pick its battles and leave posts uncontested that it would have liked to intervene on. Second, the candidate articles that were considered as potential targets of an intervention are known and retrievable from the moderators' internal chat log. This enables me to compare actual interventions to a control group of "runner-up" posts. These are articles that the group considered as targets for an

[14]The group was founded in late December 2016. By the time I started the data collection it already had 35,000 members, a number that grew to 37,000 by the end of the observation period.

[15]As the team of volunteers did not have automated tools at the time, there is no sharp discontinuities that I could exploit for identification.

[16]The call to action comes with a link to the article and with a link to a tool that helps identify other group members to facilitate targeted "liking" of other group members' comments. The group has a common hashtag that makes it easy to recognize these comments.

intervention but ultimately did not intervene on. I argue that between comparable candidate articles among which the moderators had to choose, the assignment of the intervention is as-if random.

Specifically, I construct the control group as follows. From the internal chat protocol of the group's moderators in charge of selecting the posts for intervention I retrieve all links that were discussed in the group. This set of links contains all the posts that were targeted by an intervention as well as all candidate posts that were considered for intervention. Of course, not at every point in time were there multiple posts that were equally likely to become the target of an intervention. Sometimes, there was just one article that clearly attracted most hateful comments and no comparable runner-up. The set of candidate posts therefore needs to be restricted in order to construct a comparable control group. I run a logistic regression to predict intervention probabilities for each post from the chat by the log number of comments and likes and the share of hateful comments over a 100 and a 30 minutes interval before intervention, as well as the stock of those numbers 100 minutes from intervention.[17] For each intervention post, I then retain only the three closest potential control posts that fall within plus or minus 5 percentage points distance in terms of the predicted intervention probability.[18]

In Appendix 1.C.4 I report detailed robustness checks for alternative ways of constructing treatment and control group from the chat log. In particular, I test a LASSO to predict intervention probabilities, retain all treatment posts, vary the number of retained control posts and add additional temporal restrictions on the matches. I find that the key results presented in the remainder of the paper are robust to making these modifications.

Table 1.2: Funnel of potential treatment and control posts

	Treatment	Control	Total
All posts in sample			249,426
All posts in chat log	315	1,727	2,042
Retained posts	178	370	548

Note: Summary of the number of potential treatment and control posts. The first row reports the total number of media posts in the sample. The second row contains all those posts that are mentioned in the chat log of the counterspeech group's leaders. They are divided into posts on which an intervention took place (treatment) and the rest (control). The last row reports the number of posts that are retained using the restrictions explained in Section 1.3.

The procedure results in a set of 178 treatment posts and 370 control posts, thereby leveraging 57% of the potential treatment posts mentioned in the chat log (Table 1.2).[19] Treatment

[17] Defining these time intervals for potential control posts requires making an assumption about when the counterfactual intervention would have happened. Here I assume the same time interval from the publication of the article to the intervention as the treatment article. I discuss this and possible alternatives in more detail in Section 1.5.

[18] Retaining up to three control posts for each intervention post strikes a good balance between similarity of the posts on one hand and statistical power on the other, but the results presented here are mostly robust to keeping only the closest, or the five closest potential control posts.

[19] The other interventions are not used for identification because they do not have sufficiently similar articles from the potential control group that could be used as counterfactual and were not already matched to another treatment post. In Appendix 1.C.4, I report a robustness check in which I retain all treatment posts and obtain

and control posts are published by similar media pages and pertain to similar topics. Table 1.3 breaks down the articles in treatment and control group by their topic category. As expected, the majority of articles in both groups talk about refugees or politics. The "miscellaneous" category is a frequent target of hate speech when articles about crimes break, leading users to speculate about the perpetrators' origins.[20] Table 1.16 in the appendix shows that posts in the treatment and control group were published by the same news media and in comparable proportions.

Table 1.3: Posts by post's topic in treatment and control group

	Treatment		Control	
	No.	%	No.	%
Business	2	1.1%	6	1.62%
Miscellaneous	25	14.0%	69	18.65%
Other	21	11.8%	62	16.76%
Politics	33	18.5%	56	15.14%
Refugees / Foreigners	97	54.5%	175	47.30%
Sports	0	0.0%	2	0.54%
Total	178	100.0%	370	100.00%

Note: Number and column percentages of treatment and control articles by topic category. "Miscellaneous" contains articles about accidents, crimes and disasters. "Other" is a catch-all category containing articles about science, human interest stories, celebrity news, cooking, etc. The "weather" category is empty and therefore not reported.

Prior to intervention, treatment and control group are comparable in a number of key observables. Table 1.4 reports summary statistics of treatment and control posts before the start of the intervention and shows no significant differences with the exception of the number of comments which is slightly lower in the control group. However, the magnitude of this difference compared to the total number of comments and reactions is relatively small. Compared to the full sample of posts collected over the observation period, treatment and control posts attracted a lot more activity and received a much higher share of xenophobic comments and likes.

Figure 1.5 shows that both groups' pretrends in terms of overall activity are closely aligned before intervention. Once an intervention starts, the number of comments and likes per minute almost immediately and persistently diverge.

In addition to a set of treatment and control *posts*, the identification strategy can also be used to assign treatment and control status to *individuals*. Treated individuals are those who commented on an article or liked another user's comment on an article prior to an intervention on that article. Conversely, we can assign users to the control group if they were active on a control post and thus narrowly "escaped" treatment. This assignment rules out that users self-selected into or out of treatment as it would have been very difficult for them to predict whether the counterspeech group would intervene on the treatment or the control post.

qualitatively similar results.

[20] While the two articles from the sports section may seem surprising at first, they in fact report incidences with extreme right wing fans during matches of the German national football team.

Table 1.4: Posts by post's topic in treatment and control group

	Treatment	Control	Δ	Full sample
Comments	99.2	85.8	-13.4*	16.4
Reactions	399.2	395.9	-3.3	79.7
Users	253.9	255.6	1.7	57.9
Xen. comments (%)	28.1	27.6	-0.5	3.3
Comments with tags only (%)	0.9	0.8	-0.1	2.8
Observations	178	370	548	249,425

Note: The columns report pre-intervention information on the posts in the treatment group, the control group, the difference between the two and the full set of all posts in the sample. The first row reports the average number of comments written on a post prior to intervention, the second one row contains all likes and other reactions to user comments. The third row corresponds to the average number of users. Row four reports the share of comments and likes that are xenophobic. Row five reports the share of comments who only tag or reference another user. This is often done to attract attention of specific users to an article. All values are computed just before an intervention is announced, except for the full sample, where this time is undefined. Instead the values in the last column are computed 52 minutes after a post's publication, which corresponds to the median time between publication and intervention.

Table 1.5: Observables on treatment and control users

	Treatment	Control	Δ	Full sample
Avg. weekly comments, likes	5.06	4.94	-0.12	1.26
Avg. weekly hateful comments, likes	0.40	0.40	0.00	0.07
Share of weeks w. activity	0.70	0.69	-0.01	0.3
Share of weeks w. hateful activity	0.25	0.23	-0.02	0.01
# of commented media outlets	4.30	4.23	-0.07**	2.27
Observations	20,342	61,508	81,850	1,654,729

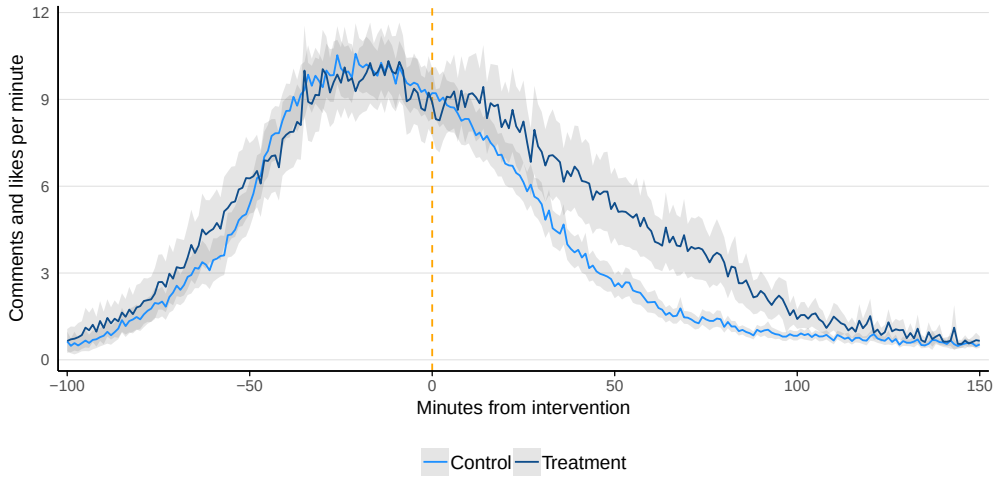
Note: The columns report the average pre-intervention activity of users in the treatment and control group, as well as the difference between the two. For comparison, the last contains averages for the full sample of users with activity on at least two days.

Individuals in treatment and control group are indeed almost indistinguishable in terms of observable behavior prior to treatment. Table 1.5 compares both groups of users and finds no statistically significant differences in terms of weekly activity levels or share of hateful comments. The only difference seems to be in the number of media outlets a user actively commented on: treatment users are active on slightly more media pages than control users.[21]

Note that the users in treatment and control group are not the average German Facebook user and not even the average user commenting the news on Facebook. Compared to the full sample of individuals in the database, these users are much more likely to be in the right tail of overall activity levels, with an average number of weekly comments and likes of about 5, whereas the average for users who appear at least twice in my sample is 1.3. The fact that both groups of users are so similar despite the fact that treatment and control group are selected solely on post-level characteristics gives additional assurance that they can be used for identification of the treatment effects of the interventions.

[21] I did not collect demographic information on users and hence cannot provide comparisons along these dimensions.

Figure 1.5: Evolution of the number of comments and likes per minute in treatment and control group



Note: Number of comments and reactions per minute on treatment and control posts by minutes to the announcement of an intervention by the counterspeech group. The shaded gray areas correspond to 95% confidence intervals.

1.4 Impact on individuals' future behavior

1.4.1 Impact on propensity to engage in hate speech

Are individuals who were subject to a counterspeech intervention less likely to write or condone hateful comments in the future? To answer this question I employ a differences-in-differences strategy using the treatment and control groups of users described in the previous section: I compare users who experienced a treatment event, i.e. users who commented or liked a comment on a post that was targeted by an intervention before its start, to users who experienced a control event, i.e. users who commented or liked a comment on a control post.

Since even the users in the treatment and control sample only write or like an average of 0.4 hateful comments per week, I aggregate the data to weekly time intervals for each user. Moreover, in most specifications I will focus on the binary dependent variable if a user wrote or liked a hateful comment in a given week, rather than studying the number of hateful comments which is skewed and contains many zeros.[22]

In order to verify that the two groups have similar activity patterns prior to the treatment or control event, I plot the event-study graph of the intervention using the following linear probability model:

$$HateSpeech_{it} = \sum_{\tau=-5}^{5} \delta_\tau \times \mathbb{1}\{t = \tau\} \times Treatment_{ie} + \alpha_i + \beta_\tau + \gamma_t + \varepsilon_{it} \tag{1.1}$$

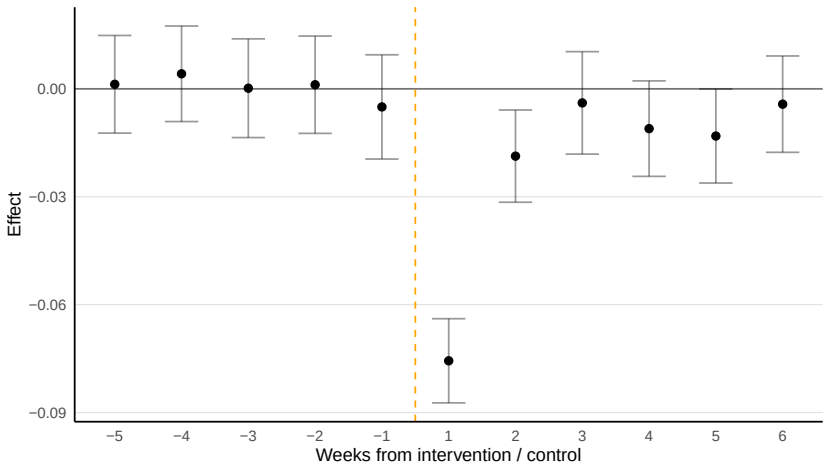
Observations are indexed by individual i, the relative time to and from treatment τ during each ten-week window e constructed around treatment and control events, and the calender

[22]The results are robust to the alternative approach of using a Poisson regression, as reported in Table 1.17 in the appendix.

week t. The subscript e is needed because individuals can witness multiple treatment and control events, a fact that I will investigate further below. The dependent variable is an indicator equal to one if the individual writes or likes a hateful comment in a given week, $Treatment$ indicates whether the event falls into the treatment or control category. α_i, β_τ, and γ_t are individual fixed effects, time relative to the event dummies, and week fixed effects respectively.

Figure 1.6 reports the coefficients δ_τ of this regression along with 95% confidence intervals based on standard errors clustered at the user level.[23] In the weeks prior to an event, both users who experience a control event and those experiencing a treatment event are equally likely to write or condone a hateful comment in a given week, the coefficients δ_τ for $\tau = 1, ..., 5$ being small and insignificant. Activity patterns quickly diverge upon treatment: the week after being exposed to an intervention, treated users are about 7.6 percentage points less likely to engage in hate speech than the comparison group. The effect persists less strongly for another week, at 1.9 percentage points, before it decays. There still seem to be slight effects four to five weeks after treatment but they are only weakly significant and less robust to alternative ways of clustering the standard errors.

Figure 1.6: Intervention impact on individuals' propensity to write or like xenophobic comments



Note: This event-study plot graphs the δ_τ coefficients from regression (1.1) along with 95% confidence intervals based on user-clustered standard errors. It corresponds to the second column of Table 1.17 in the appendix, which also contains robustness checks using logistic and Poisson regressions.

These results imply that the counterspeech interventions are highly effective at deterring users from engaging in hate speech – for a short period of time. Compared to the treatment users' baseline probability to write or like a hateful comment of 0.25, the effects' magnitude is sizable. Note that this temporary change in behavior is similar to the findings of Munger (2017) who reports that calling out users that post racist tweets reduces their likelihood of repeating this behavior for about a month before the effect wears off.

[23]The standard errors are somewhat larger when clustered at the event level, but the drop at week 1 and 2 remains statistically significant at conventional levels.

1.4.2 Heterogeneity in interventions' effectiveness

In order to investigate the drivers of this effect and to see which users respond most to the interventions, I collapse the two weeks before and after an event into a pre- and a post-period and perform a standard differences-in-differences analysis. Column 1 and 2 of Table 1.6 confirm the results of the event-study: In the two weeks after the intervention, treated users are on average 5.3 percentage points less likely to engage in hate speech. This result is robust to controlling for the share of hateful comments on the event post, and for an individual's average number of comments and likes per week prior to the intervention, both hateful and not. This confirms that the drop in the probability of spreading xenophobic content is not driven by individuals in treatment and control group engaging in discussions at different frequencies already before the intervention.

Table 1.6: Differences-in-differences regression results at the user level

	$\mathbb{1}\{Xen.\ Comment/like\}$				
Intervention	-0.053^{***}	-0.052^{***}	0.010	-0.049^{***}	-0.049^{***}
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)
× hates ≤ weekly			-0.139^{***}		
			(0.009)		
× hates > weekly			-0.031^{**}		
			(0.011)		
× small Intervention				−0.007	
				(0.010)	
× large Intervention				−0.008	
				(0.010)	
× share xen. comments on article					-0.559^{***}
					(0.031)
Controls		Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes
Users	101,274	101,274	101,274	101,274	101,274
Observations	248,219	248,219	248,219	248,219	248,219
R^2	0.620	0.620	0.621	0.620	0.622

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the forth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The effect seems to be driven mainly by users who do engage in hateful activity, but do so only occasionally. Column 3 of Table 1.6 disaggregates the treatment effect for users who have not previously written or liked a hateful message, users who did do so but less than weekly, and users who engage in this kind of activity more than once a week. This grouping of users approximately corresponds to tertiles of hateful activity. Witnessing a counterspeech intervention has no effect on users who do not engage in hate speech in the first place. At 13.9 percentage points, the effect is strongest for users who write or like hateful comments at less

than weekly frequency, i.e. moderately hateful users. Conversely, the treatment is much less effective on extreme users who very frequently engage in hateful activity.

I next turn to the question if the interventions' effectiveness depends on the relative importance of hate and counterspeech on the intervention post. The answer is not clearcut. When interacting the intervention with its size as measured by tertiles of the share of comments on the intervention post that were written by the counterspeech group, the bulk of the effect seems to emanate from medium sized interventions (column 4 of Table 1.6). Both small and large interventions, on the other hand, seem to have no discernible effect. It is important to note, however, that the level of participation in the interventions is likely to be endogenous. For instance, it is conceivable that users who were active on posts which attracted interventions with overwhelming support are more extreme users who could be less likely to respond to the treatment.[24] The share of xenophobic comments on the treatment post, on the other hand, is highly negatively correlated with the size of the effect. The more hate speech was pronounced on the intervention post, the higher the subsequent drop in an individual's propensity to engage in hate speech (last column). Again, it is important to stress that the variation in the share of xenophobic comments is not random and already observable by individuals at the time they decide to participate in the discussion.

The extent to which an individual responds to an intervention depends on whether that individual has been singled out during the intervention *specifically*. Interventions consist of members of the group writing counterspeech messages as comments on the targeted article generally (top-level comments) but also of commenting directly to the hateful messages that were made by other users (sub-level comments). This allows to look at the treatment effect for those individuals who received a public reply to their comment by a member of the counterspeech group. This is true for about 54% of the users who were active before an intervention occurred. Table 1.7 introduces an interaction term capturing this situation into the differences-in-differences regressions. While the effect of experiencing an intervention remains negative and statistically significant, its magnitude is much smaller than in the previous regressions. Individuals who received a direct counterspeech reply, however, are an additional seven percentage points less likely to engage in hate speech over the weeks after the intervention. As before, users who only occasionally write or condone xenophobic comments are affected the most by an intervention and there is little additional effect of direct replies for more hateful individuals. How many replies a user receives does not seem to alter the effectiveness of the intervention. In conjunction with the fact that I found no effect of the size of the intervention, this result suggests that the key driver of the effects here is being targeted individually by a another user.

These results are unlikely to be accounted for by deletions of comments or accounts. One could be concerned that participants in the counterspeech interventions report users who engage

[24] I considered several potential instruments for the size of the interventions: First, I computed the number of the counterspeech group's members that participated in the previous days or week with the idea that a larger group should be able to stage bigger interventions. Second, I tried to identify times of the day on different weekdays during which participation in the interventions would increase, hoping to find a lunch- or coffee break effect. Unfortunately, none of these proved to be sufficiently predictive to be used in an IV strategy.

Table 1.7: Differences-in-differences regression with narrow treatment definition

	$\mathbb{1}\{Xen.\ Comment/like\}$			
Intervention	−0.016*	−0.016*	−0.016*	−0.016*
	(0.007)	(0.007)	(0.007)	(0.007)
Intervention × direct reply (SLC)	−0.071***	−0.070***	−0.019	−0.078***
	(0.009)	(0.009)	(0.011)	(0.011)
× hates ≤ weekly			−0.120***	
			(0.012)	
× hates > weekly			−0.001	
			(0.015)	
× log(SLC)				0.006
				(0.005)
Controls		Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
Users	101,274	101,274	101,274	101,274
Observations	248,219	248,219	248,219	248,219
R^2	0.620	0.620	0.621	0.620

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

in hate speech to Facebook which in turn suspends or deletes these individuals. However, I confirm in Appendix 1.C.3 that the results also hold for the subsample of individuals remaining active in the weeks following treatment or control events.

1.4.3 Impact on individuals' activity

How do counterspeech interventions impact individuals' activity patterns more broadly? Rather than just moderating the content of their comments, users who are affected by an intervention seem to reduce their activity in general. Table 1.8 contains the results of applying the differences-in-differences estimation to activity as measured by the total number of comments and likes, both hateful and not. The first column shows that treated users become 5.7 percentage points less likely to be active in a given week, which is quite close to the magnitudes of the effect on hate speech. There also seems to be a small but insignificant intensive margin effect on activity, as suggested by column 2. Combining intensive and extensive margin suggests an overall reduction of 6.8% in the number of comments and likes following an intervention.[25]

Similarly to the effects on hate speech, the reduction in activity seems to predominantly come from moderately hateful users that have written hateful messages before, but write them

[25] I use $log(1+\#comments\ and\ likes)$ as a dependent variable to measure the combined effect on intensive and extensive margin. In Appendix 1.C.2 I show that the results are broadly robust to using a Poisson regression which may be statistically more appropriate but more difficult to interpret in particular in the presence of high-dimensional fixed effects.

less than once a week. While there still is no evidence for the effects increasing in the size of the intervention, the reduction in activity is driven by the upper two tertiles of interventions, with small interventions not having a significant impact. In contrast to the results on the propensity to engage in xenophobic comments, the interaction of the intervention with the share of hate speech on the targeted article is small and statistically insignificant.

Table 1.8: Differences-in-differences on user activity

	Measure of activity					
	$activity > 0$	$log(activity)$	$log(1 + activity)$			
Intervention	-0.057^{***}	-0.018	-0.068^{***}	-0.015	-0.059^{***}	-0.067^{***}
	(0.003)	(0.009)	(0.008)	(0.011)	(0.010)	(0.008)
× hates ≤ weekly				-0.115^{***}		
				(0.015)		
× hates > weekly				-0.035		
				(0.022)		
× small Intervention					0.006	
					(0.016)	
× large Intervention					-0.056^{**}	
					(0.019)	
× share xen. comms.						-0.092
						(0.054)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	101,274	85,994	101,274	101,274	101,274	101,274
Observations	248,219	213,937	248,219	248,219	248,219	248,219
R^2	0.501	0.801	0.810	0.810	0.810	0.810

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the fifth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The decrease in activity is not driven solely by the reduction in hate speech. Instead, treated individuals reduce both xenophobic *and* other activity. Table 1.18 in the appendix presents the results of repeating the differences-in-differences analysis using only non-xenophobic activity as the dependent variable. The magnitudes of the effects are slightly smaller than in the previous table, but the patterns in terms of heterogeneity remain broadly the same.

In addition to being less likely to write or like comments in general and xenophobic comments in particular, I find evidence that individuals who were exposed to a counterspeech intervention also shift their remaining activity towards less contentious topics. In order to see whether the interventions render individuals more or less likely to express themselves on certain matters, I disaggregate the data such that I can track each users' activity by topic category of the articles they discuss. I can thus treat each individual-topic-week triple as a separate observation. I

then run the following differences-in-differences regression:

$$Active_{itce} = \delta_{TREAT} \times Treatment_{itce} + \delta_{CONTR} \times Control_{itce} + \\ + \sum_{\varsigma \in C} \delta_{\varsigma} \times Treatment_{ite} \times \mathbb{1}\{c = \varsigma\} + \alpha_{ie} + \beta_{ic} + \gamma_{w} + \varepsilon_{itce} \tag{1.2}$$

Here, subscript i denotes individuals, t indicates the two-week period before or after the event, c denotes the topic category in the set of categories C, and e denotes the position in the sequence of events that the individual is exposed to. The covariates are a dummy indicating whether the individual experienced a treatment event on on the given topic, a dummy indicating whether the individual experienced a control event on that topic, and a dummy whether the individual experienced a treatment on *any* topic, interacted with the topic category at hand. The δ_{ς} coefficients capture how individuals' activity on each topic is affected by an intervention. In addition, the regression controls for individual-event fixed effects, user-topic, and week dummies. The standard errors are clustered within users.

Table 1.9: Substitution patterns

	Measure of activity		
	$\mathbb{1}\{act > 0\}$	log(act)	log(1+act)
Intervention on topic	−0.279***	−0.174***	−0.409***
	(0.005)	(0.012)	(0.007)
Control on topic	−0.175***	−0.065***	−0.301***
	(0.002)	(0.005)	(0.003)
Int. on other topic × biz. econ.	0.003	0.059***	0.029***
	(0.003)	(0.015)	(0.004)
Int. on other topic × miscellaneous	−0.028***	−0.0003	−0.042***
	(0.004)	(0.011)	(0.005)
Int. on other topic × other	−0.021***	−0.060***	−0.089***
	(0.003)	(0.009)	(0.005)
Int. on other topic × politics	−0.008*	−0.024*	−0.065***
	(0.003)	(0.010)	(0.006)
Int. on other topic × refugees	−0.080***	−0.001	−0.105***
	(0.004)	(0.013)	(0.007)
Int. on other topic × sports	0.019***	0.078***	0.054***
	(0.003)	(0.017)	(0.004)
Int. on other topic × weather	0.029***	0.129***	0.066***
	(0.002)	(0.028)	(0.003)
Controls	Yes	Yes	Yes
User × event FE	Yes	Yes	Yes
User × topic FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	1,737,533	740,247	1,737,533
R^2	0.779	0.852	0.854

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The excluded topic category is "other" articles, which contain celebrity news, lifestyle articles, movie reviews, etc. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The estimates for the δ_{ς} coefficients are reported in Table 1.9. Both treatment and control

events are followed by a drop in activity in the topic category in which the event occurred. That activity decreases for both types of events may be explained by individuals having satisfied their desire to express their opinion on a given topic. Consistent with the overall treatment effect, however, the decrease is more pronounced for treatment than for control events. The spillovers of interventions to other topic categories reveal a clear pattern: the effects are negative for topic categories prone to heated discussions and higher shares of hate speech, such as immigration, politics more broadly, and miscellaneous. The exception is formed by articles in the "other" category, for which activity drops as well. Less contentious subjects, such as business, sports, and the weather, see an increase in activity. Taken together, this suggests that interventions induce individuals to withdraw from the debates that are most prone to xenophobia and instead engage in the discussion of topics where hateful comments are more rare.

1.4.4 Multiple treatments

Individuals can be subject to multiple treatment and control events thus allowing for an analysis of how the magnitudes of the effects vary with the number of treatments. Whereas 74% of users in the sample are exposed to only one event during the observation period, there is still a sizable number of individuals who experience multiple events. However, the sample size quickly decreases in the number of prior events.[26]

The effect sizes could change with the number of treatments due to two possible mechanisms. The first one is a change in the effectiveness of the treatment itself. For instance, the treatment effect could wear off as users learn about a social norm or internalize information, as I will discuss further in in Section 1.6. The second mechanism could be selection: As I showed in the previous section, users respond to treatments by being less likely to write a hateful comment or indeed any comment at all for a period of time. As a consequence, there is a form of survivor bias in the sense that individuals who are still actively commenting after a treatment and are thus available to be treated another time are precisely those users who were less responsive to the first treatment.

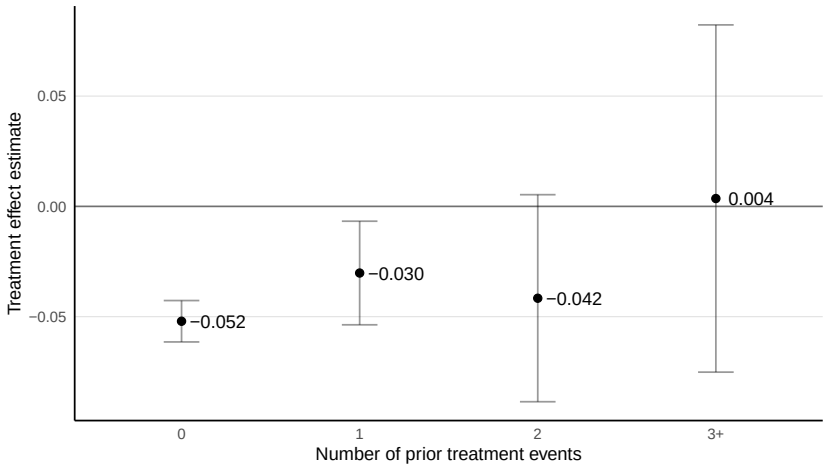
While I cannot disentangle the two mechanisms, the identification strategy explained above allows to measure their combined effect. Since it implies that assignment to treatment or control is as good as random for each event, I can compare the response to treatment of individuals who have the same number of prior treatment events. I thus estimate the treatment effect conditional on remaining active after a given number of treatments rather than the average effect of receiving a given number of treatments. I do so by interacting treatment status with the number of previous exposures to treatment and estimating the following regression:

[26] Figure 1.11 in the appendix shows the number of event and user tuples by the number of times a given user has been previously exposed to treatment and control events at the time of the treatment or control event.

$$HateSpeech_{ite} = \sum_{p=0}^{P} \delta_p Treatment_{ite} \times \mathbb{1}\{PriorTreatments = p\}_{ie} + \\ + \sum_{p=0}^{P} \nu_p \mathbb{1}\{PriorControls = p\}_{ie} + \alpha_i + \gamma_t + \varepsilon_{it} \tag{1.3}$$

Here, subscript i denotes individuals, t indicates the two-week period before or after the event, and e denotes the position in the sequence of events that the individual experiences. The relevant parameters here are the set of δ_p which estimate the effect of interventions conditional on a users prior history of treatments.

Figure 1.7: Diff-in-Diff estimate on xenophobic comments & likes by treatment history



Note: Regression coefficients δ_p from equation 1.3 with 95% confidence intervals based on standard errors clustered at the user level. The regression results are also reported in more detail in Table 1.19 in the appendix.

Figure 1.7 plots these estimates for users who had no prior exposure to treatment during the observation period, one prior treatment event, two prior treatment events, and three or more prior treatment events. As the number of prior treatment events increases, the effect of each additional treatment decreases. The effect size of the first treatment has roughly the size of the average treatment effect, but already the second treatment is about half as effective. Beyond that, the loss of precision and significance driven by the shrinking sample size makes it difficult to say exactly how large the treatment effects still are. However, there is a clear pattern in the point estimates suggesting that there is no additional treatment effect beyond the second treatment event.

1.4.5 Substitution towards more extreme pages

The fact that treated individuals are less likely to engage in hate speech and to comment on the public Facebook pages of news media naturally leads to the question of whether they actually reduce their levels of hate speech or if they just voice their opinions elsewhere, for instance among more like-minded people, in private comments or on fringe outlets. If we care about the size of the audience that xenophobic ideas are able to reach, then the fact that interventions

reduce the amount of hate speech on articles of large news media is already very good news. However, one might also be concerned that some individuals could progressively be pushed towards a more radical fringe.

Since the scope of my data only includes the public pages of large German language news media, I cannot speak to what these individuals do or say in private Facebook groups, on other websites, or even offline. I can, however, leverage the fact that even within the 85 outlets in my data, there is considerable heterogeneity in terms of the size of the audience, the editorial mix, the prevalence of xenophobic comments and the likelihood of being subject to an intervention (see Table 1.15 in the appendix). If individuals radicalized as a result of a counterspeech intervention, we could expect to observe a shift of their activity towards outlets that attract more like-minded users. Even if these are not completely radical fringe outlets, they would probably be more appealing to progressively radicalizing individuals than news media with more mainstream commentators.

To investigate if there is empirical support for this radicalization hypothesis I disaggregate the data such that each observation is identified by a triple of individual i, media page p, and two week period t before or after event e . Similar to the analysis of spillovers between topics, I then estimate the following differences-in-differences regression:

$$\begin{aligned} Active_{itpe} = \delta_{TREAT} \times Treatment_{itpe} + \delta_{CONTR} \times Control_{itpe} + \\ + \delta \times Treatment_{ite} + \alpha_{ie} + \beta_{ip} + \gamma_t + \varepsilon_{itp} \end{aligned} \tag{1.4}$$

The dependent variable is a dummy indicating if an individual liked or wrote a comment on a given page. The covariates are a dummy indicating whether the individual experienced a treatment event on the given page, a dummy indicating whether the individual experienced a control event on that page and a dummy whether the individual experienced a treatment on *any* of page, as well as individual-event fixed effects, user-page, fixed effects and time dummies.

I find little evidence in the data that would suggest that individuals switch to pages with higher levels of hate speech or a lower probability of being targeted. The first column of Table 1.10 contains the results of estimating regression 1.4. The effect of an intervention is not just contained to the page on which the intervention took place, but spills over to other pages as well. The second column shows that the effect seems to be larger for pages with a bigger audience. This is presumably because users are more likely to be active on these pages in the first place, which is why I retain this interaction term as a control in the remaining regressions.[27] In the third column, I introduce an interaction with the total number of counterspeech interventions on a page. While the point estimates of the effects are largest for pages that were especially often targeted by interventions, the differences between the coefficients are very small compared to the average magnitude of the effect: a 1.3 percentage point difference between pages that never see an intervention versus those who do so most frequently. The last column shows that after an intervention, users decrease their activity most on pages that have a high share of

[27]Figure 1.12 in the appendix shows the spillover-effect for each page in the sample for completeness.

Table 1.10: Impact of interventions on activity by pages

	$\mathbb{1}\{act > 0\}$			
Intervention on page	-0.259^{***}	-0.254^{***}	-0.249^{***}	-0.254^{***}
	(0.004)	(0.004)	(0.004)	(0.004)
Control on page	-0.157^{***}	-0.157^{***}	-0.157^{***}	-0.157^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Intervention on other page	-0.0004^{***}	0.029^{***}	0.010^{***}	0.029^{***}
	(0.0001)	(0.001)	(0.001)	(0.001)
× log(# page followers)		-0.003^{***}	-0.001^{***}	-0.002^{***}
		(0.0001)	(0.0001)	(0.0001)
× 1-30 treatments			-0.007^{***}	
			(0.0005)	
× >30 treatments			-0.013^{***}	
			(0.001)	
× page share of xen. comments				-0.047^{***}
				(0.005)
User × event FE	Yes	Yes	Yes	Yes
User × page group FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	21,098,615	21,098,615	21,098,615	21,098,615
R^2	0.844	0.844	0.844	0.844

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Page followers" is the number of users who liked a news media page. "1-30 treatments" and ">30 treatments" are dummy variables indicating that a media page was targeted by the corresponding number of interventions during the observation period. The omitted category is pages that were never targeted by an intervention during the observation period. "Page share of xen. comments" is the average share of xenophobic comments and likes on a given media page.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

xenophobic comments. The magnitude of this interaction is again very small, but lends little support to the idea that individuals radicalize in response to the intervention.

1.5 Immediate impact on targeted articles

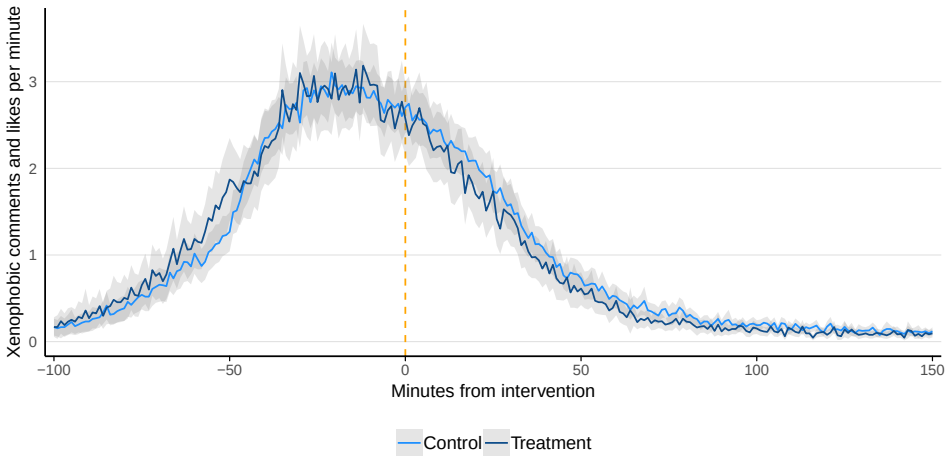
The identification strategy outlined in Section 1.3 can also be used to asses the more immediate impact of interventions on the targeted articles' discussions. The analyses presented in the previous section addressed the question of how the future behavior of individuals already actively participating in a debate responds to an intervention. This section complements these findings by shedding light on users' aggregate response to a given article, including those individuals that were not yet participating in the discussion before the intervention. I will show that counterspeech interventions do not seem to significantly decrease the total *number* of hateful comments on the targeted articles, but do attract more moderate users beyond the circles of the intervention group to the debate, thereby leading to an overall decrease of the *share* of hateful comments.

This can be achieved by comparing articles that were targeted by an intervention to the set of control articles that were considered for intervention by the group and performing event-study and differences-in-differences analyses. However, the timing of the counterfactual interventions

poses an additional challenge that needs to be addressed for this purpose. The responses to articles on Facebook follow a highly cyclical pattern: when an article is posted, it quickly reaches an attention peak, usually within the first 30 minutes before activity slowly wears off over the subsequent couple of hours. As a result, identification of the impact of the intervention requires not only assuming *on which posts* the counterfactual interventions would have taken place, but also *when* they would have taken place.

A naive counterfactual timing would consist of attributing the actual time of intervention to the counterfactual intervention. However, since articles are posted at different times by different news outlets, this would lead to confounding treatment and life-cycle effects. Instead, I measure the time it took between the article's publication and the actual intervention and apply the same time difference to the control posts: if an intervention was announced t minutes after the targeted article was posted on Facebook, then I assume that the counterfactual interventions would have occurred t minutes after the publication of the control posts. An alternative strategy that leads to broadly similar results would have been to predict the time of intervention based on observables of the article such as the media outlet, the number of comments and likes and the share of hateful comments. Empirically that strategy leads to less well-aligned pretrends which is why I focus on the simpler method here.[28]

Figure 1.8: Xenophobic comments and likes per minute in treatment and control group



Note: Number of xenophobic comments and reactions per minute on treatment and control posts by minutes to the announcement of an intervention by the counterspeech group. The shaded gray areas correspond to 95% confidence intervals.

Figure 1.8 compares the average number of comments and likes per minute that contain xenophobic messages on treatment and control posts. The two curves track each other closely prior to the intervention lending credibility to the identifying assumptions. However, contrary to the patterns in terms of the total number of comments and likes documented in Figure 1.5, there seems to be no sharp divergence following the intervention.

[28] Yet another alternative is to use the time of discussion among the counterspeech group's moderators instead of the articles' publishing time as reference point. The pretrends seem to align less well with this method, however.

In order to test this more throughly, I estimate the following event-study model:

$$Y_{pt} = \sum_{\tau=-T}^{T} \delta_\tau \times \mathbb{1}\{t = \tau\} \times Treatment_p + \alpha_p + \beta_t + \varepsilon_{pt} \tag{1.5}$$

Where Y_{pt} is an characteristic of article p at t minutes after treatment, for instance the average number of comments per minute. α_p and β_t are article and time-since-announcement fixed effects. The coefficients of interest are δ_τ which will capture the difference between treatment and control group. For $\tau \leq 0$, δ_τ is expected to be indistinguishable from zero if treatment and control posts are truly comparable. If that is the case, we should be able to interpret δ_τ with $\tau > 0$ as the causal effect of the intervention on Y.

As a first step, I confirm that the announcements of interventions by the counterspeech group did trigger higher activity levels. Figure 1.9a plots the δ_τ coefficients of the event-study regression with the log of the number of comments and likes as a dependent variable. Consistent with the identifying assumptions, the point estimates are small and statistically insignificant prior to treatment. Within ten minutes after the interventions were announced by the group's moderators, there is an increase in the activity levels which becomes statistically significant at conventional levels 20 minutes after treatment. At its peak, the effect corresponds to a 50% increase in the number of comments and likes.

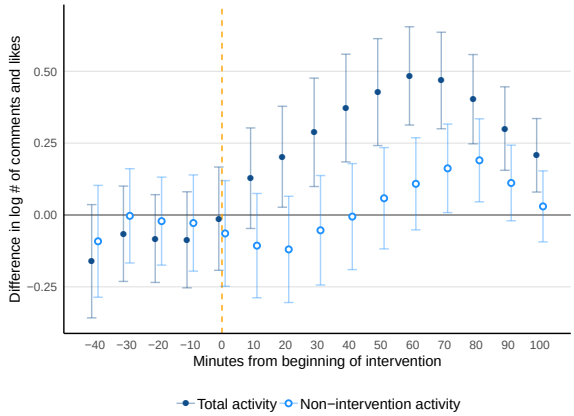
Interestingly, the increase in activity following the intervention does not seem to be entirely explained by the members of the counterspeech group carrying out the intervention. A ripple-on effect becomes apparent when estimating the event-study regression with non-intervention activity as a dependent variable, i.e. comments and likes by users who are not members of the counterspeech group and did not participate in interventions before. Figure 1.9a shows that there is an increase in activity not directly attributable to the group after about an hour after the intervention starts. Further inspection reported in Appendix 1.C.5 indicates that this is driven by an influx of users who were not active on the post before, rather than by the initial commentators replying to the intervention messages.

An event-study regression on the log number of xenophobic comments and likes suggests a slight decrease in response to an intervention by up to to 17% after about an hour after the treatment starts (Figure 1.9b). This result should not be over-interpreted, however, as it is relatively sensitive to the exact timing of the counterfactual intervention. Still, it is clear that there is no *increase* in the absolute number of hateful comments and likes on the targeted posts, which is remarkable given the fact that there is such a strong increase in activity and the number of users on the article that are not directly attributable to the counterspeech group. The first two columns of Table 1.11 confirm this pattern in a classical differences-in-differences regression which aggregates the data into a pre- and a postintervention period. Averaged over the entire post-period, the results suggest a 14% decline in the number of hateful comments and likes, with the previous caveat still applying.

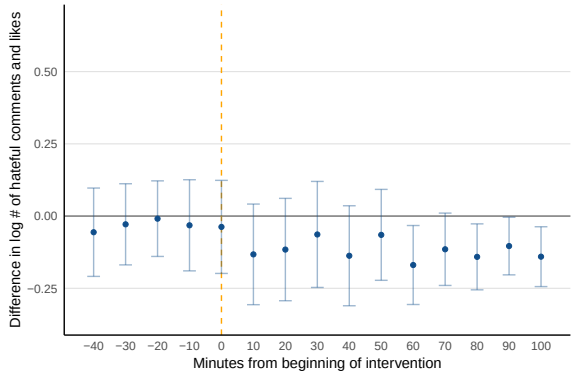
As a result, the share of xenophobic activity in the comments and likes that were not written

Figure 1.9: Event-study graphs at the article level

(a) All Comments & likes



(b) Xenophobic comments & likes



Note: Event-study graph corresponding to regression (1.5) with 95% confidence intervals based on post-clustered standard errors. Regressions include time-trends and post dummies. Non-intervention activity includes all comments and likes which have been written by users who did not write a counterspeech message in the ongoing or a previous intervention.

by users active in the counterspeech group declines. Columns three and four of Table 1.11 show that this net share of hateful messages decreases by up to 3% when averaged over the entire post-intervention period. Compared to an average share of 21.2% of xenophobic messages on these posts, this effect is sizable but far from fully eradicating hate speech in this context.

Are the additional users attracted by the intervention less likely to engage in hate speech because they are more moderate in general or because they adapt their behavior depending on the comments that they find on the article? The last two columns of Table 1.11 point to the former explanation. I compute users' average share of hate speech in all comments they made before commenting on the intervention article for all users that have not previously participated in an intervention. Using this share as a dependent variable in the differences-in-differences regression, I find that the composition of users changes in favor of ex ante more moderate users once the intervention is under way.

These results suggest that the interventions have a mobilizing effect on bystanders that lead to overall moderation of the content written in response to the articles. The intervention leads to new users joining the discussion of the targeted article who are less likely to post hateful

Table 1.11: Diff-in-diff regression results at the post level

	log(1+# xen. coms)		Share xen. (excl.CS)		User share xen.	
Intervention	−0.106	−0.136*	−0.028**	−0.030**	−0.010***	−0.010***
	(0.066)	(0.064)	(0.011)	(0.011)	(0.002)	(0.002)
Post FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes		Yes		Yes	
Time × page FE		Yes		Yes		Yes
Posts	548	548	548	548	548	548
Observations	20,824	20,824	14,490	14,490	14,349	14,349
R^2	0.594	0.616	0.236	0.274	0.193	0.226

Note: Post-clustered standard errors in parentheses. Share xen. is the number of xenophobic comments and likes which have been written by users who did not write a counterspeech message in the ongoing or a previous intervention divided by the total number of comments and likes by these users. User share xen. is the average share of xenophobic comments made in all previous comments by users who did not write a counterspeech message in the ongoing or a previous intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

messages than the initial set of users who were responding to the article. The results also imply that the sharp drop in the likelihood of an individual writing or condoning a hateful message observed in the previous section is not driven by a moderating effect on the intervention article itself. Rather, individuals who are subject to an intervention change their behavior going forward, even if its only for the limited timespan of about two weeks.

The fact that the share of xenophobic comments on an article decreases as a result of the intervention is encouraging. Ultimately, this share may be the more relevant metric than the absolute number of hateful comments: Especially when discussions contain lots of comments, a lower share of xenophobia implies a lower probability that an onlooker would be exposed to these ideas while browsing through the comments.

1.6 Possible mechanisms

Through which channels do the interventions affect individuals' behavior? Plausible candidates for mechanisms include information provision, individuals learning about social norms through the distribution of others' actions, and fostering social norm compliance through non-monetary punishment. I discuss each of these possible mechanisms in turn.

1.6.1 Information provision

The simplest way to conceptualize a counterspeech intervention would be to think of it as providing information. Interventions could convey two different kinds of information. First, counterspeech messages could reveal new information on the contentious subject at hand. If initially xenophobic individuals learned that their statements were based on erroneous beliefs about the world, for instance a much larger number of refugees coming to Europe than there actually is, then an intervention could be effective simply by correcting those beliefs. Second,

in a very different vein, targeted individuals could potentially infer from the outraged reactions of others that there may be a risk of having their account blocked or deleted.

The fact that the effect of the counterspeech interventions wears off for individuals who experience multiple such interventions could be interpreted as evidence consistent with this mechanism. If the first time an individual is confronted with counterspeech may still learn new information, it may have already absorbed this information the second time it is exposed to it.

While I cannot rule out that this channel plays a role, I argue that it is unlikely to explain the full extent of the effects observed here. As noted in Section 1.2.3, only 14% of the counterspeech messages in response to the article contain factual information on the topic such as statistics or facts relevant to the article and most of these are actually drawn from the article itself. This share is even lower, about 4%, for responses to other users' comments made by participants in the counterspeech intervention. Yet it is precisely those responses that seem to drive the largest effect.

The treatment effect's decay over a few weeks is also difficult to reconcile with information provision. We would expect either that individuals are able to retain the information learned from an intervention, or that an additional intervention refreshes their memory. Yet, neither of these patterns emerges from the analyses presented above.[29]

It is similarly unlikely that the counterspeech interventions led targeted individuals to update their beliefs about the risk of being blocked or deleted from specific media pages or from Facebook in general. First, I show in Appendix 1.C.3 that deletions themselves cannot explain the drop in activity, which means that there would be no factual basis to sustain such a belief. Second, there are barely any threats or reminders in the counterspeech group's comments. In only 1% of the direct responses to other users' comments did I find a mention of reporting the user, and there is even fewer in the comments on the article.

1.6.2 Inference of social norms from average behavior

A more plausible mechanism is that the interventions induce individuals to update their beliefs about a prevailing social norm. Counterspeech could shift individuals' perceptions of what type of messages the average user does or does not condone. If individuals derive utility from writing comments that do not deviate too much from average opinion, then this change in perception could lead them to adapt their conduct. This type of behavior has been documented for instance by Bursztyn et al. (2017), who show that respondents are more willing to express racist views once they have learned that more people agree with these statements than they previously thought.

This mechanism would be consistent with the fact that the overarching theme which all the counterspeech messages have in common is indeed to express disagreement with hateful views,

[29]Moreover, it is not clear whether correcting users' knowledge would be sufficient to alter their behavior so dramatically. Using a survey experiment in a similar context, Barrera Rodriguez et al. (2018) find that fact-checking changes the beliefs about the state of the world of participants who were previously exposed to false claims, but does not alter the conclusions they draw from these false claims.

be it by explaining the reasons why these views are unacceptable, by making injunctive norm statements or by just providing their own opinion. However, this channel too has difficulty explaining the fact that while being exposed to multiple interventions progressively reduces the treatment effect, the effect of any intervention also decays over time. If individuals inferred a social norm from the behavior of others, then these two findings would be hard to reconcile.

It would also imply that there should be a correlation between the share of counterspeech messages on the targeted article and the effectiveness of the intervention. If counterspeech makes up the vast majority of the comments, then this should shift beliefs more than if there is only a few scattered messages, all else equal. Yet, as I show in Section 1.4, I do not find such a correlation in the data. The reason for this might be that the size of the interventions is endogenous to the article topic. It could be conceivable that participation in the interventions is higher for topics with a broader consensus. Individuals who still post hateful messages on these topics are maybe more extreme types that would be less responsive to treatment, irrespective of how many people participate in the interventions. In the absence of a good exogenous shifter of participation it is impossible to conclude on whether this specific type of social norms channel would be consistent with the data.

1.6.3 Social norm enforcement through non-monetary punishment

While both of the aforementioned mechanisms could play a role in explaining the observed effects, the impact on individual behavior seems most closely aligned with the findings of the experimental literature on non-monetary punishment. In the seminal paper of this literature, Masclet et al. (2003) study a repeated public goods experiment in which participants were able to attribute non-monetary "punishment points" to other members of their group. They found that despite these points having no effect on participants' payoffs, subjects used these points to punish low contributors which in turn responded by increasing their contributions. The effectiveness of non-monetary punishment has been replicated in a series of studies and notably with written messages that participants were able to send to other group members.[30]

Analogously, counterspeech can be thought of as a form of non-monetary punishment. Members of the intervening group write messages in which they publicly disapprove of the behavior of individuals who write hateful messages. The results presented here are notably consistent with the findings of the non-pecuniary punishment literature in three important aspects. First, I find that the effects of the intervention on individual behavior are primarily attributable to counterspeech comments written directly in response to a user's own comment (see Table 1.7). What seems to matter is the fact of being individually targeted by an expression of disapproval. Second, the decay of the interventions' effect over time resembles the results typically obtained in experimental public goods games with non-monetary punishment. A closer look at the results obtained by Masclet et al. (2003) for instance reveals that while the option to punish increases contributions rates on average, it does not seem to remedy the decay of cooperation

[30]See for instance Ellingsen and Johannesson (2008), Xiao and Houser (2009) and Dugar (2010).

over the course of many rounds of the repeated game. Quite to the contrary, decay might even be steeper. Finally, as noted before, the counterspeech messages consist mainly of expressions of disagreement with and disapproval of the hateful messages that were written by other users and are therefore in line with the idea of punishment. In this respect, their content is quite comparable to the messages that the participants in laboratory experiments wrote in response to other participants' contribution decisions.[31]

Whether non-monetary counterspeech sanctions are effective because their public nature induces shame in the targeted users or because they raise the awareness and salience of an existing social norm cannot be decided based on the findings presented here. While there is a longstanding literature arguing for the effectiveness of shame caused by public sanctions (see for instance Kahan and Posner (1999)), Xiao and Houser (2011) have shown that penalties are more effective when given in public than in private even when anonymity rules out shame as a mechanism. In fact, the findings on non-monetary sanctions described above are obtained entirely in settings that exclude shame as a possible explanation.[32] Rather than by a behavioral response to avoid a negative emotion such as shame, these results can be explained by the sanctions communicating a social norm (Xiao and Houser (2011)) or by raising its salience (Konow (2000) and Xiao and Houser (2009)). The fact that counterspeech is most effective when directed specifically at individual users is consistent both with inducing shame and with communicating the norm most effectively, as replies are almost sure to be read by the targeted individuals. In lack of any variation of the publicity of the sanction, I can cannot determine which of the two mechanisms is at play.

1.7 Conclusion

In the face of a rising tide of populism and xenophobia, people have often pointed fingers at social media in the popular press and the broader public debate. As evidence of the sometimes dramatic real world consequences of online hate speech is becoming increasingly available, it has become more important than ever to understand the drivers that shape this behavior.

In this paper, I show that individuals' desire to behave in ways that will be judged favorably by others is one of these drivers. I demonstrate that German users of the world's largest social media platform, Facebook, reduce their supply of online hate speech in response to organized "counterspeech" interventions by fellow users. To do so, I collected roughly 12 million comments on the public Facebook pages of German language news media and identify hateful comments using recent deep learning techniques. Identification is obtained by comparing the treatment group of news articles and of users that were targeted by an intervention to a plausible control group that was equally likely to be targeted. The latter group is inferred from the internal chat

[31] See Table 3 of Xiao and Houser (2009). For example: "[...]you need to check your priorities...it's not about the money, it's about sharing what you have and realizing you're not the center of the world" (p.399).

[32] Smith et al. (2002) and others have argued that shame requires some degree of publicity of the transgression. In Masclet et al. (2003), however, actions were communicated without any identifying information, not even anonymized player numbers or IDs.

logs of the organizers of these large scale counterspeech interventions.

Comparing targeted individuals to the control group I find that the treated individuals are substantially less likely to engage in counterspeech over a period of two weeks. Rather than moderating the opinions they express, these users are less likely to voice their beliefs altogether and tend to avoid contentious debates. While a number of possible mechanisms could explain the effectiveness of these interventions, the findings seem to be most plausibly explained by the interventions acting as a form of non-monetary punishment that makes individuals fall back in line with a prevailing social norm of acceptable behavior.

I also find that the interventions have a moderating effect on the discussion of the articles that were targeted by the intervention. While not reducing the overall number of hateful comments made on these articles significantly, interventions do attract users to the discussion that express non-hateful views, thereby reducing the share of hateful comments.

If confirmed, the findings presented in this paper carry both good news and bad. The good news is possibly that users of social media do adjust their behavior in response to others, demonstrating that maintaining a dialog even in the context of sometimes toxic online debates has clear benefits. The bad news may be that the mechanism I describe may just as well work the other way around: unchecked hate speech may set off a vicious cycle triggering more hate speech. In this context, the fact that media companies seem to have an incentive to slant news towards articles that trigger hateful reactions in order to gain attention on social media is particularly worrying and highlights the special responsibility that editorial boards need to exercise.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *Quarterly Journal of Economics:* 1885–1939.

Barrera Rodriguez, Oscar David, Sergei M. Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2018. "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics." *Working paper.*

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652–1678.

———. 2012. "Laws and Norms." *IZA Discussion Paper* 6290:1–44.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *arXiv,* no. 1607.04606.

Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao. 2017. "Status Goods: Experimental Evidence from Platinum Credit Cards."

Bursztyn, Leonardo, and Robert Jensen. 2016. "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure." *NBER Working Paper.*

Dugar, Subhasish. 2010. "Nonmonetary sanctions and rewards in an experimental coordination game." *Journal of Economic Behavior and Organization* 73 (3): 377–386.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of lncentive Theory." *American Economic Review* 98 (3): 990–1008.

Enikolopov, Ruben, Alexey Makarin, Maria Petrova, and Leonid Polishchuk. 2017. "Social Image, Networks, and Protest Participation."

Gächter, Simon, and Ernst Fehr. 1999. "Collective action as a social exchange." *Journal of Economic Behavior & Organization* 39 (4): 341–369.

Gagliardone, Iginio, Gal, Danit, Alves, Thiago, Martinez, and Gabriela. 2015. *Countering Online Hate Speech.* Paris: UNESCO.

Gentzkow, Matthew, Bryan T. Kelly, and Matt. Taddy. 2017. "Text as Data." *Working Paper.*

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78 (1): 35–71.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." *NBER Working Paper:* 46.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120 (4): 1191–1237.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach." *The Quarterly Journal of Economics* 133 (2): 801–870.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780.

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2013. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity* 2012 (1): 1–81.

Kahan, Dan M., and Eric A. Posner. 1999. "Shaming White-Collar Criminals: A Proposal for Reform of the Federal Sentencing Guidelines." *Journal of Law and Economics* 42 (S1): 365–392.

Kennedy, Patrick J, and Andrea Prat. 2019. "Where do people get their news?" *Economic Policy* 34 (97): 5–47.

Konow, James. 2000. "Fair Shares : Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90 (4): 1072–1091.

Lee, Dong-Hyun. 2013. "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks." *working paper.*

Loughran, Tim, and Bill Mcdonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (4): 1187–1230.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review* 93 (1): 366–380.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv,* no. 1301.3781: 1–12.

Müller, Karsten, and Carlo Schwarz. 2018a. "Fanning the Flames of Hate: Social Media and Hate Crime."

———. 2018b. "Making America Hate Again? Twitter and Hate Crime Under Trump."

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39 (3): 629–649.

Noussair, Charles, and Steven Tucker. 2005. "Combining monetary and social sanctions to promote cooperation." *Economic Inquiry* 43 (3): 649–660.

Reuters. 2018. *Why Facebook is losing the war on hate speech in Myanmar,* August.

———. 2019. *Germany fines Facebook for under-reporting complaints.*

Siegel, Alexandra A. 2018. "Online Hate Speech."

Smith, Richard H., J. Matthew Webster, W. Gerrod Parrott, and Heidi L. Eyre. 2002. "The role of public exposure in moral and nonmoral shame and guilt." *Journal of Personality and Social Psychology* 83 (1): 138–159.

Xiao, Erte, and Daniel Houser. 2009. "Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange." *Journal of Economic Psychology* 30 (3): 393–404.

———. 2011. "Punish in public." *Journal of Public Economics* 95 (7-8): 1006–1017.

Yanagizawa-Drott, David. 2014. "Propaganda and Conflict: Theory and Evidence From the Rwandan Genocide." *Quarterly Journal of Economics* 129 (4): 1947–1994.

Ziegele, Marc, Teresa K Naab, and Pablo Jost. 2019. "Lonely together? Identifying the determinants of collective corrective action against uncivil comments." *New Media & Society.*

Appendices

1.A Machine classification

The algorithm used to categorize users' comments into hateful and non-hateful proceeds is a sequence classification model that proceeds in two steps. First, using an unsupervised approach, the text gets a representation as a sequence of vectors. Then, this sequence is fed into Long Short-Term Memory (LSTM) and a feed-forward neural network that predicts the most likely class.

1.A.1 From comments to sequences of vectors

A comment can be interpreted as a sequence of vectors, where each vector represents one token in the comment. A token can be a word, an emoticon (e.g. ":)") or punctuation. As a first step, each token is represented by a m-dimensional dummy-vector e_i with the i-th component equal to one and all other components equal to zero, where m is the number of unique tokens in the corpus of comments. Using these representations to make predictions directly, is complicated by the fact that the vectors are very large, sparse and that their relative position in the vector space has little meaning.

To reduce the dimensionality of these vectors and to get more meaningful token representations, I use the Word2Vec algorithm proposed by Mikolov et al. (2013). The algorithm consists of a neural network with a single hidden layer of l neurons that is trained to predict a token e_i in a comment based on the surrounding tokens in the same comment. The l weights thus obtained for each token are normalized to unit length and used as new l-dimensional representation w_i of the token.

In addition to having lower dimensionality and being dense, these vector representations have a key feature: The cosine distance between them can be used as a measure of semantic similarity. Intuitively, this comes from the fact that two words that are surrounded by the same set of words often have similar meaning. As the same set of words is predictive for both of them, their vector representations will be similar as well. This is very useful in the context human language where the same idea can be phrased in a myriad of ways, only a fraction of which will be observed in a relatively small training sample. For example, the word "Asylant" (asylum seeker) and "Flüchtling" (refugee) describe similar concepts, yet this relation is absent in their dummy vector representations e_i, e_j. Their representation w_i, w_j, however, based on the data have cosine a cosine similarity $w_i \cdot w_j = 0.84$. In practice this helps the classifier categorize "Asylanten raus" and "Flüchtlinge raus" the same way. Since these representations are learned in an unsupervised manner, they can be learned on a larger training corpus than the hand-classified training corpus so that the classifier can make predictions on comments

consisting of words that never occurred in the training sample.

To obtain 300-dimensional token representations, I train the Word2Vec algorithm on all 18.8 million user comments and comments on comments written by users, the 249,000 media posts, as well as the full text of more than 200,000 news articles referenced int the posts. I exclude tokens that occur less than 50 times and use a context window of ten tokens to predict each token. For illustration, I list the nearest neighbors for a few tokens in table 1.12. They show that vector representations help deal with common typos and misspellings ("Flüchtling" vs. "Flüchling"), that they capture related concepts (Merkel is the last name of the German chancelor, "kanzlerin") and that they can help identify racist slurs ("Goldstücke", "Kulturbereicherer").

Even using a large training corpus, some words (or indeed misspellings of words) occur too rarely to compute token vectors, despite the fact that they do occur in the comments. Germans seem to be particularly prone to using composite nouns, concatenating several (frequently occurring) nouns to build a new (rarely occurring) noun. To deal with this issue, if a token is encountered that does not have a token vector associated with it, I compute a new token vector by averaging over all known tokens that I can find within the unknown token.[33]

Table 1.12: Examples of tokens with similar vector representations

flüchtling	flüchling	merkel	:)	goldstücke
asylant	asylant	murksel	;)	goldjungs
migrant	flüchtling	merkels	:-)	bereicherer
wirtschaftsflüchtling	wirtschaftsflüchtling	mekel	:p	neubürger
kriegsflüchtling	kriegsflüchtling	merkl	;-)	kulturbereicherer
schutzsuchender	wirtschaftsmigrant	kanzlerin	:slightly_smiling_face:	goldstückchen

Note: This table reports the five closest tokens of five example tokens base on the cosine distance between their vector representations obtained by Word2Vec. The English translations are "refugee","refgee" (typo), "merkel" (the last name of the German chancelor Angela Merkel), an emoticon and "piece of gold" - a term that is used as a slur by some users.

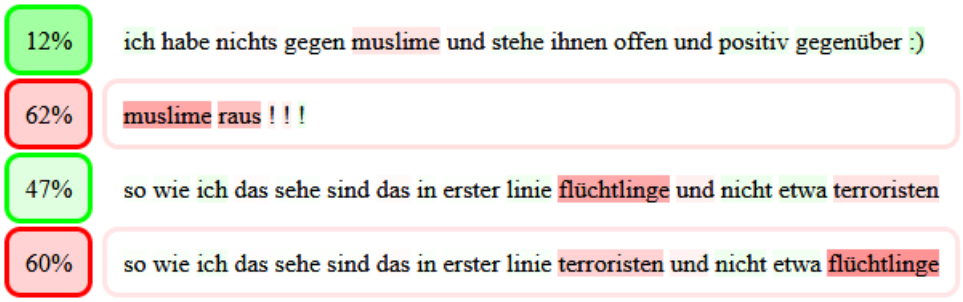
1.A.2 From token vectors to comment classification

The token vectors computed in step one are then used to turn each comment into a sequence of token vectors that are fed into a LSTM network for prediction. The LSTM is a particular form of a neural network that has been shown to deal well with long-term dependencies (Hochreiter and Schmidhuber (1997)). It reads the token vectors one-by-one and updates its cell state at after each token is read in. How much each token affects the cell state is dependent on the token and all tokens previously seen by the LSTM. The final cell state at the end of a comment is a meaning encoding of that comment. In order to allow the model to also take into account tokens that

[33] A similar idea is used by the FastText algorithm (Bojanowski et al. (2017)) that uses within-word n-grams to enrich word vectors. In my tests, however, FastText was outperformed by the approach outlined above, even when using only syllable n-grams.

LSTM-generated encodings have a number of useful properties which are illustrated using a few example comments in figure 1.10 which shows the predicted probability that a comment contains hateful content. First, the same token can affect the cell state differently depending on the tokens that preceded in the sequence. The first two examples show that the word "Muslime" (muslims) increase the likelihood that a comment is classified as hateful, but less so in the first example, where it is used in a positive context. Second, they depend on the order of tokens, which avoids ambiguity. Examples three and four contain exactly the same tokens and differ only in their order, which completely changes the meaning of the sentence. In this instance, the classifier correctly classifies both of them. This property would be difficult to achieve using a bag-of-words approach without resorting to very long n-grams that further exacerbate the sparsity problem.

Figure 1.10: Classification examples using LSTM



Note: The left column shows the probability that a comment should be classified as xenophobic predicted by the classifier. Each word is colored by the extent to which the predicted probability would be higher (red) or lower (green) if that word was removed from the comment. The comments are inspired by actual comments found in the data but have been modified for illustrative purposes. They translate to : "I have nothing against muslims and I have an open and positive attitude towards them", "Muslims out!!!", "The way I see it these are first and foremost refugees and by no means terrorists", "The way I see it these are first and foremost terrorist and by no means refugees".

The encodings computed by the LSTM are then passed into a fully connected feed-forward neural network with. At the last step, a normalized exponential function (softmax) is applied so that the network returns a probability that a given comment is hateful. The model is then trained to minimize the negative log-likelihood.

1.A.3 Parametrization and performance of the classifier

The model that achieved the best performance consists of two consecutive bidirectional LSTM layers, each with a 50-dimensional output space, and dense layer with 64 neurons. The input sequences are truncated to a maximum of 50 tokens. A ten percent dropout is applied to avoid overfitting in the training phase. Once the model is trained, predictions are made for 5000 thousand unlabeled comments that are not part of the training data. Comments that are assigned with at least 65 percent confidence to either class are added to the training sample along with their probabilities and the model is trained again on the augmented data. This pseudo-labeling approach inspired by Lee (2013) is repeated for two iterations and slightly

improves performance. The final classifiers' average performance in a ten fold cross-validation is reported in table 1.13.

Table 1.13: Classifier's average performance in ten fold cross-validation

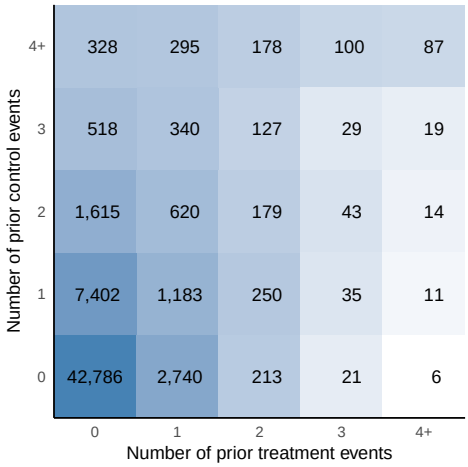
Performance measure	Classifier performance
Accuracy	94.4%
Area under ROC	93.4%
Specificity	97.3%
Sensitivity	56.9%

Note: All performance measures are computed on the hold-out set in 10 fold cross-validation. Accuracy is the share of correct predictions. The area under the receiver operating characteristic curve is the integral over the curve plotting the true-positive rate against the false-positive rate for each probability threshold of predicting positive outcome. Specificity is defined as the share of correctly predicted negatives in all negatives in the sample. Sensitivity is the share of correctly predicted positives in all positives in the sample.

I find that the exact architecture and parametrization of the model matters relatively little for the predictive power of the model, but that the most important driver of performance are the token embeddings. In a way, the part of the model that is decribed in step two seems to be the proverbial tip of the iceberg which accounts for "only" about 200,000 thousand parameters of the model while the token vectors account for roughly 25 million (approx. 84,000 tokens). What seems to matter most for performance is (1) the fact that the network is recurrent as opposed to using a purely constitutional neural network for instance and (2) that it uses domain-specific token-vectors created directly from task-related text as opposed to widely used net-crawls and Wikipedia dumps.

1.B Additional tables and figures

Figure 1.11: Number of user-event tuples by prior treatment exposure



Note: Number of user × treatment/control-event cells broken down by the number of prior treatment and control events that a user has experienced. The top row and rightmost column contain cells with four and more prior treatment events and control events respectively.

Table 1.14: Full list of news media included in the data with number of followers and media posts.

Facebook Page	Followers	Posts
Bild	2394020	1188
SPIEGEL ONLINE	1439586	4248
tagesschau	1348413	2167
RTL Aktuell	1110008	1250
N24	1063394	3174
WELT	920983	2271
n-tv Der Nachrichtensender	841519	1699
ZEIT ONLINE	823817	5441
stern	726010	6319
ZDF heute	703057	4879
Süddeutsche Zeitung	689118	820
FOCUS Online	683579	1439
HuffPost Deutschland	644071	5252
FOCUS Online Politik	531325	989
FAZ.NET - Frankfurter Allgemeine Zeitung	486204	1107
Süddeutsche Zeitung Magazin	475061	185
DIE ZEIT	423062	530
DW (Deutsch)	389498	5369
BILD News	388444	3641
Zeit im Bild	365929	399
derStandard.at	296121	3453
RT Deutsch	288129	3641
Kronen Zeitung	269631	5014
taz. die tageszeitung	268264	801
WELT Video	253376	2587
Handelsblatt	213681	1029
Berliner Zeitung	179554	1244
nrw-aktuell.tv	171853	1371
Aktuelle Stunde	170882	1284
DiePresse.com	163588	2210
Deutschlandfunk	161246	1805
ZDF heuteplus	154607	2351
NZZ Neue Zürcher Zeitung	152981	2268
Kleine Zeitung	137012	2515
BILD am SONNTAG	134502	2304
MDR - Mitteldeutscher Rundfunk	103755	2619
Deutsche Wirtschafts Nachrichten	100588	1083
BILD Hamburg	100507	1940
Nürnberger Nachrichten	100460	1186
shz.de - Nachrichten aus Schleswig-Holstein	95570	1416
SWR Aktuell	94627	1469
Mitteldeutsche Zeitung	94426	676
MDR Sachsen-Anhalt	92871	3319
Notruf	88805	612
Passauer Neue Presse - PNP	86417	5257
Ostsee-Zeitung	86009	841
stuttgarter-nachrichten.de	85695	2460
Frankfurter Rundschau	85325	5719
Islamische Zeitung	79899	415
Hannoversche Allgemeine Zeitung / HAZ	79684	2078
stuttgarter-zeitung.de	79536	6090
Badische Zeitung	77900	1369
hessenschau.de	77751	3970
SAT.1 Nachrichten	75797	1705
Nürnberger Zeitung	75601	8662
nachrichten.at – Oberösterreichische Nachrichten	73767	4418
Süddeutsche Zeitung München	72771	2300
Lübecker Nachrichten Online	68862	4266
Westfälische Nachrichten	67752	2244
Polizei Nachrichten Österreich	65345	2578
Ruhr Nachrichten	62447	882
svz.de - Nachrichten aus Mecklenburg-Vorpommern	58809	7647
rbb24	55648	6437
MDR Sachsen	50694	9458
Allgemeine Zeitung	50018	3944
BILD Dresden	48887	3241
Brandenburg aktuell	43692	7176
MDR Thüringen	42277	6463
BILD Leipzig	40655	563
SÜDWEST PRESSE Online	39273	575
BILD Köln	38512	9103
FINANCIAL TIMES DEUTSCHLAND	36811	1372
BILD Frankfurt	35793	5993
Wiener Zeitung	35165	613
BILD Saarland	29010	1798
BILD Bremen	26550	188
BILD Politik	22890	473
BILD Hannover	22809	476
BILD Ruhrgebiet	20777	1328
BILD Düsseldorf	18064	4703
BILD Berlin	16034	5001
BILD München	14763	7401
BILD Thüringen	11234	5319
BILD Chemnitz	9348	4027
BILD Stuttgart	8907	996
Total		250113

Note: This table lists all 85 Facebook pages that were monitored from August 2017 through January 2018. The number of followers was recorded in July 2017. It includes most major German news outlets along with some smaller regional outlets. A few fringe pages were included as well.

Table 1.15: Posts, user activity and hateful comments by news media

		Articles		Users	
	Total activity	# Articles	% Xen.	# Users	% Xen.
Bild	5,322,821	8,662	4.4	1,243,642	5.4
Spiegel Online	2,411,631	6,464	2.6	353,357	6.3
N24	2,195,287	5,318	8.8	283,451	12.4
Focus Online	2,102,473	9,447	7.8	168,066	18.5
Tagesschau	2,062,436	3,316	3.6	311,122	8.3
Focus Online Politik	1,974,441	5,994	10.0	97,311	28.2
Rt Deutsch	1,823,520	9,104	9.7	140,440	27.3
Welt	1,590,224	4,027	5.0	224,288	9.0
Zdf Heute	1,386,921	3,174	9.2	161,187	14.0
Rtl Aktuell	1,342,266	5,437	7.0	250,383	11.1
Huffpost Deutschland	1,263,817	6,439	10.2	95,350	22.4
Kronen Zeitung	1,114,944	4,988	10.5	78,828	21.9
Süddeutsche Zeitung	1,020,689	5,721	2.7	154,490	6.2
Frankfurter Allgemeine Zeitung	966,820	7,648	5.5	86,767	12.8
N-TV	888,942	4,877	7.2	115,299	16.4
Other	9,508,964	158,810	3.8	2,058,838	7.1

Note: Number of posts, user reactions and the xenophobic share thereof for the top 15 media pages by user activity. The columns report the total number of comments and likes per media outlet, the total number of articles, the average share of xenophobic activity, the total number of users and the share of users who wrote or condoned a xenophobic comment at least once. Note that the fact that the tabloid Bild has such a low share of xenophobic content is driven by the fact that many of its posts are celebrity news. Where they do post topics that are more prone to hate comments, this share is higher.

Table 1.16: Posts by media page in treatment and control group

	Treatment		Control	
	No.	%	No.	%
Zdf Heute	39	21.9%	54	14.6%
N24	22	12.4%	51	13.8%
Bild	18	10.1%	41	11.1%
Tagesschau	20	11.2%	37	10.0%
Huffpost Deutschland	11	6.2%	26	7.0%
Spiegel Online	10	5.6%	25	6.8%
Focus Online	10	5.6%	24	6.5%
Focus Online Politik	7	3.9%	24	6.5%
Kronen Zeitung	9	5.1%	22	5.9%
Rtl Aktuell	11	6.2%	17	4.6%
Other	21	11.8%	49	13.2%
Total	178	100.0%	370	100.0%

Note: Number and column percentages of treatment and control articles by media outlet. Only the 10 most frequently targeted media outlets are reported.

Table 1.17: Intervention impact on individuals' propensity to write or like xenophobic comments

δ_τ	OLS		Logit		Poisson	
$\tau = -5$	0.001	0.001	0.033	0.068	0.076	0.079
	(0.007)	(0.007)	(0.038)	(0.067)	(0.055)	(0.055)
$\tau = -4$	0.004	0.004	0.068	0.112	0.051	0.053
	(0.007)	(0.007)	(0.036)	(0.062)	(0.051)	(0.050)
$\tau = -3$	0.0003	0.0002	0.041	0.085	−0.030	−0.027
	(0.007)	(0.007)	(0.037)	(0.064)	(0.049)	(0.048)
$\tau = -2$	0.002	0.001	0.039	0.062	−0.036	−0.034
	(0.007)	(0.007)	(0.035)	(0.061)	(0.044)	(0.044)
$\tau = -1$	−0.005	−0.005	−0.057	−0.099	−0.048	−0.045
	(0.007)	(0.007)	(0.034)	(0.061)	(0.032)	(0.031)
$\tau = 1$	−0.075***	−0.076***	−0.466***	−0.818***	−0.488***	−0.487***
	(0.006)	(0.006)	(0.029)	(0.051)	(0.039)	(0.039)
$\tau = 2$	−0.018**	−0.019**	−0.094**	−0.162**	−0.117**	−0.118**
	(0.007)	(0.007)	(0.031)	(0.053)	(0.039)	(0.039)
$\tau = 3$	−0.004	−0.004	0.008	0.004	−0.004	−0.009
	(0.007)	(0.007)	(0.037)	(0.063)	(0.044)	(0.045)
$\tau = 4$	−0.011	−0.011	−0.006	−0.005	0.011	0.010
	(0.007)	(0.007)	(0.035)	(0.061)	(0.048)	(0.048)
$\tau = 5$	−0.013	−0.013*	−0.037	−0.060	−0.045	−0.041
	(0.007)	(0.007)	(0.035)	(0.061)	(0.051)	(0.051)
$\tau = 6$	−0.004	−0.004	0.030	0.051	−0.004	−0.002
	(0.007)	(0.007)	(0.035)	(0.060)	(0.048)	(0.048)
Controls		Yes		Yes		Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Time to event FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	426,979	426,979	221,325	221,325	227,204	227,20

Note: This table reports the results of different specifications for the event-study regression described in Section 1.4. The coefficients in the second column are plotted in Figure 1.6. The first two columns contain the linear probability model described in equation 1.1. The middle two columns report the results of logistic regressions with the same dependent variable. The last two columns contain the results of Poisson regressions where the dependent variable is the number of xenophobic comments written or liked by an individual in a given week. User-clustered standard errors in parentheses. The number of observations excludes singletons and, for the GLMs, perfect classifications. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.18: Diff-in-Diff estimate on non-xenophobic comments & likes

	Measure of non-xenophobic activity					
	$activity > 0$	$log(activity)$		$log(1 + activity)$		
Intervention	−0.052***	0.002	−0.046***	0.004	−0.036***	−0.047***
	(0.003)	(0.010)	(0.008)	(0.011)	(0.010)	(0.008)
× hates ≤ weekly				−0.107***		
				(0.015)		
× hates > weekly				−0.041		
				(0.022)		
× small Intervention					0.009	
					(0.016)	
× large Intervention					−0.065***	
					(0.019)	
× share xen. comms						0.157**
						(0.054)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	101,281	99,953	101,281	101,283	101,283	101,282
Observations	248,219	224,415	248,219	248,219	248,219	248,219
R^2	0.534	0.811	0.807	0.808	0.807	0.807

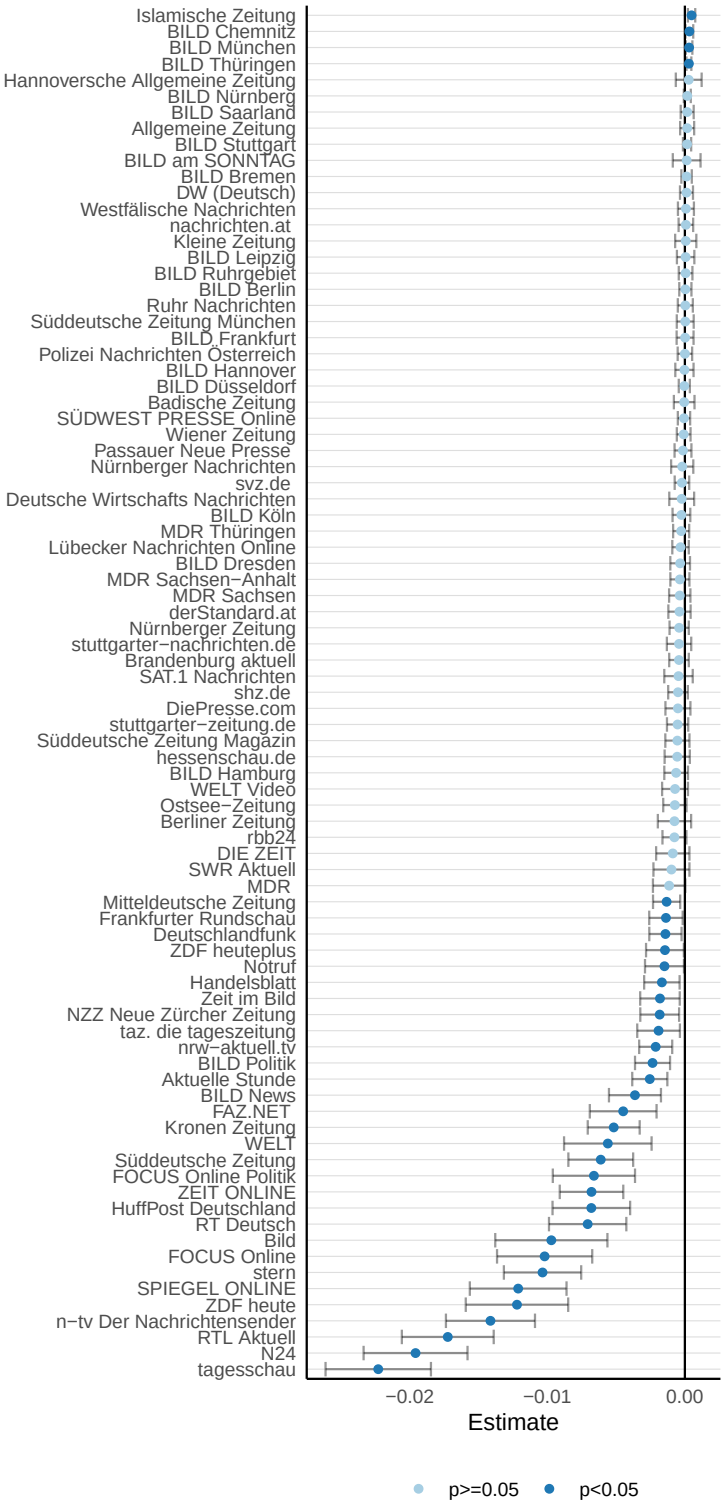
Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.19: Diff-in-Diff estimate on xenophobic comments & likes by treatment history

	$\mathbb{1}\{Xen.\ Comment/like\}$	
No previous treatment	−0.051***	−0.052***
	(0.005)	(0.005)
1 previous treatment	−0.035**	−0.030*
	(0.012)	(0.012)
2 previous treatments	−0.045	−0.042
	(0.024)	(0.024)
≥3 previous treatments	−0.003	0.004
	(0.040)	(0.040)
Controls		Yes
Control event FE	Yes	Yes
User FE	Yes	Yes
Period FE	Yes	Yes
Users	101,274	101,274
Observations	248,219	248,219
R^2	0.620	0.620

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The coefficients in the second column are plotted in Figure 1.7. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.12: Spillover-effect between media pages



Note: This figure reports the impact a counterspeech intervention on a users probability to write or like a comment on a given Facebook page. It plots the δ coefficient of the differences-in-differences regression $Active_{itp} = \delta_{treat} \times Treat_{itpe} + \delta_{contr} \times Contr_{itpe} + \delta \times Treat_{ite} \times I_p + \alpha_{ie} + \beta_{ip} + \gamma_t + \varepsilon_{itp}$, where I_p is a page dummy. Standard errors are clustered at the user level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.C Additional results and robustness checks

1.C.1 Article topics and user responses

Table 1.20 reports the results of regressions predicting the share of hateful comments, the log number of comments and likes and the log number of active users on topic dummies and a media outlet × date fixed effect. Each regression is performed on the full set of pages as well as on the core sample of media outlets, i.e. those pages on which the counterspeech group at least considered to intervene at some point during the observation period.

Table 1.20: User response to articles by topic category

	Share hate coms.		log(activity)		log(#users)	
Business / Economics	0.001	−0.001	0.053	0.075*	0.034	0.049
	(0.001)	(0.001)	(0.029)	(0.038)	(0.027)	(0.035)
Miscellaneous	0.020***	0.027***	0.106***	0.131*	0.082**	0.096
	(0.002)	(0.002)	(0.029)	(0.054)	(0.028)	(0.051)
Politics	0.009***	0.011***	0.709***	0.817***	0.570***	0.654***
	(0.002)	(0.002)	(0.064)	(0.101)	(0.056)	(0.089)
Foreigners / Refugees	0.082***	0.093***	0.881***	1.033***	0.728***	0.855***
	(0.004)	(0.004)	(0.068)	(0.106)	(0.060)	(0.093)
Sports	−0.007***	−0.009***	−0.363***	−0.418***	−0.351***	−0.399***
	(0.001)	(0.002)	(0.046)	(0.083)	(0.043)	(0.077)
Weather	−0.007***	−0.010***	−0.480***	−0.700***	−0.424***	−0.618***
	(0.001)	(0.002)	(0.083)	(0.135)	(0.078)	(0.130)
Page × date FE	Yes	Yes	Yes	Yes	Yes	Yes
Pages	All	Core	All	Core	All	Core
Observations	248,913	121,377	248,913	121,377	248,913	121,377
R^2	0.144	0.147	0.412	0.318	0.419	0.328

Note: Page-clustered standard errors in parentheses. The number of observations excludes singletons. The excluded topic category is "other" articles, which contain celebrity news, lifestyle articles, movie reviews, etc. The first, third and fifth column include all pages in the regression. The second, fourth and sixth column restrict the sample to core pages which have at least one article that was considered or targeted by the counterspeech group.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results show, perhaps unsurprisingly, that articles which broadly relate to foreigners receive considerably more xenophobic comments and likes than any other topic category. More interestingly, they also receive more attention by users than any other topic category. This correlation holds even within news outlets and is therefore not driven by the composition of topics and outlets. Compared to an article in the "other" category, writing about immigration can be associated with up to twice the number of comments and likes.

Pushing this idea one step further, I regress the log number of comments and likes and the log number of users active on an article on the share of xenophobic comments and likes on the article. Table 1.21 shows that these correlations remain strongly positive and highly statistically significant even when controlling for topic × date indicators and page × date indicators. A ten percentage point increase in the share of xenophobic comments is associated with a 19-25%

increase in activity and 15-20% more users on a given posts. To the extent that the social media teams of news media try to maximize engagement with users, this suggests that they may have incentive to produce content that triggers hateful reactions.

Table 1.21: User response to articles by level of hate speech

	log(activity)		log(#users)	
Share of hateful comments	1.927***	2.511***	1.583***	2.044***
	(0.170)	(0.237)	(0.147)	(0.216)
Page × date FE	Yes	Yes	Yes	Yes
Topic × date FE	Yes	Yes	Yes	Yes
Pages	All	Core	All	Core
Observations	248,909	121,377	248,909	121,377
R^2	0.432	0.348	0.436	0.354

Note: Page-clustered standard errors in parentheses. The number of observations excludes singletons. The first and third column include all pages in the regression. The second and fourth column restrict the sample to core pages which have at least one article that was considered or targeted by the counterspeech group.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.C.2 Robustness: Poisson regression results on activity

In Section 1.4 I report results showing that individuals who were targeted by a counterspeech intervention are not only less likely to engage in hate speech for a period of time, but also are less active in general. For ease of interpretation, I used $1 + log(\#comments\ and\ likes)$ as a dependent variable. Table 1.22 shows that the key result that individuals become less active after intervention can also be obtained using a Poisson regression and is therefore not dependent on the specific functional form I imposed.

While the parameter estimates confirm the general pattern presented in the main body of the text, there are some notable differences in the heterogeneity analysis. Using the Poisson model, it does seem that there is an effect on users who write hateful messages more than once a week. Small interventions have a significant effect, while the significance on larger interventions decreases. However, the estimates should be interpreted with care as the inclusion of high-dimensional fixed effects in the regression is likely to introduce incidental parameter bias.

1.C.3 Deletions of accounts

In Section 1.4 I showed that individuals who experience a counterspeech intervention are less likely to write or like xenophobic comments in the weeks after, and to write or like comments on news articles more generally. One concern regarding these results could be that the interventions trigger Facebook to block or delete certain users and that the reduction in hate speech and activity simply reflects that these users were deleted as a result of the intervention. This could be the case for instance if the participants in the intervention in addition to writing

Table 1.22: Differences-in-differences on user activity (Poisson regression)

	Activity (# comments + # likes)				
Intervention	−0.039***	−0.081***	−0.012	−0.054***	−0.080***
	(0.010)	(0.010)	(0.019)	(0.013)	(0.010)
× hates < weekly			−0.093***		
			(0.023)		
× hates > weekly			−0.094***		
			(0.024)		
× small Intervention				−0.095***	
				(0.021)	
× large Intervention				−0.021	
				(0.024)	
× Share xen. comments					−0.345***
					(0.069)
Controls		Yes	Yes	Yes	
User FE	Yes	Yes	Yes	Yes	
Period FE	Yes	Yes	Yes	Yes	
Observations	251,355	251,355	251,355	251,355	251,355

Note: The table reports the parameter estimates of Poisson regressions along with user-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the fourth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

counterspeech messages also reported users that engage in hate speech to Facebook and the company would act on these complaints.

In order to rule out that this is the main driver of the effect, I restrict the sample to users who write or like at least one comment in the weeks following a treatment or control event and repeat the differences-in-differences analysis. These users were evidently not deleted or banned from the platform and if deletions were to explain the previous findings, we would not expect to see any impact of the interventions on this subsample.

Table 1.23 reports the result of the differences-in-differences regression on the subsample of users who remained active in the two weeks following a treatment or control event. There remains a highly significant 5 percentage points drop in the individual's probability of posting or condoning a hateful message. Compared to the baseline results presented in Table 1.6, the effect is slightly smaller but still comparable in magnitude. Deletions are are therefore unlikely to account for the observed decrease in xenophobic activity.[34]

[34]Anecdotally, there have been many complaints about Facebook's lack of responsiveness to hate speech in general and reports filed by users in particular. Recently, German authorities even decided to fine the company for failure to adequately report instances of hate speech (Reuters (2019)).

Table 1.23: Differences-in-differences result on users remaining active

	$\mathbb{1}\{Xen.\ Comment/like\}$	
Intervention	-0.050^{***}	-0.049^{***}
	(0.005)	(0.005)
Controls		Yes
User FE	Yes	Yes
Period FE	Yes	Yes
Users	85,502	85,502
Observations	210,415	210,415
R^2	0.621	0.622

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.C.4 Alternative definitions of treatment and control group

In this section I test the robustness of the results presented in the main body of the paper to alternative ways of defining treatment and control groups from the counterspeech group's internal chat log. I obtain the potential control posts by extracting all urls mentioned in the group's chat and matching them to a Facebook post using a unique post identifier contained in the links.[35] In order to identify causal effects of the counterspeech interventions, it is necessary to restrict the full set of posts that were ever discussed in the group's chat to those posts that can credibly serve as counterfactual for the interventions. Here I discuss variations in those restrictions.

Baseline procedure

In Section 1.3 I outlined a two step procedure to get to a set of treatment posts and likely runner-up posts to be used as controls. In the first step, I predict the intervention probability for all posts in the chat log using a logistic regression based on pre-intervention observables. Specifically, I include the log activity levels at 100 and 30 minutes before intervention, the log number of comments and likes over the last 30 minutes prior to the intervention decision and the share of hateful comments just before the intervention decision and 30 minutes before the intervention decision as predictors. The first column of Table 1.24 reports the coefficients of this regression. The difference in the predicted treatment probability is a measure of two posts' similarity. For each treatment post I retain only candidate control costs that are within plus or minus five percentage points of predicted intervention probability. In the second step, I further restrict the set of potential control posts by retaining only the closest three of these control posts for each treatment post.

[35] I remove posts with less than 10% xenophobic comments, as these were likely posted in the group for other reasons. The results are robust to setting this threshold to 5%.

To test the main results' sensitivity to the restrictions in each of these two steps, I modify them in turn and repeat the complete set of analyses presented above.

Estimating treatment propensities using LASSO

To modify the first step, I run a LASSO logistic regression instead of the standard logit with manually chosen predictors. I tie my hands by letting the algorithm choose from a rich set of pre-intervention observables. The regularization parameter λ is chosen to maximize the area under the receiver operating characteristic curve (AUC) in 5 fold cross-validation. The second column of Table 1.24 contains the resulting coefficients as well as the full set of possible predictors. While the results turn out to be very similar to the standard logit in terms of included regressors and predictive power, I chose the standard regression as baseline because it results in slightly better balance in the treatment and control posts and is numerically more stable.

Table 1.25 presents descriptive statistics for the baseline definition used in the main paper along with the six alternative definitions described in this section. The sample obtained by using a LASSO regression and presented in the second supercolumn is almost indistinguishable from the baseline sample. Treatment and control groups balance well both in terms of articles and users.

Figure 1.13 replicates the post-level event-study plots presented in Section 1.5 for the different definitions of treatment and control posts. Broadly, all definitions yield a similar overall pattern comparable to the baseline definition: Interventions lead to a substantial increase in the number of comments and likes on targeted posts and potentially to a slight decrease in xenophobic activity. The results obtained using the LASSO are no exception (light blue diamonds in the figure). The only difference to the baseline results is that the coefficients 40 minutes before intervention are different from zero, but this seems to stem mostly from the fact that some posts have not entered the sample yet at that point in time, rather than from actual pretrends.

The event-study plot in Figure 1.14 shows the impact of counterspeech interventions on individuals' probability to write or like a hateful comment in a given week. It replicates the results presented in Figure 1.6 in the main body of the paper. The strong drop in that probability the week following the treatment is highly significant in all alternative definitions of treatment and control group. The LASSO specification (light blue diamonds in the figure) tracks the baseline results almost exactly, produces comparable persistence, and exhibits even a slightly larger drop in xenophobic activity.

Finally, Table 1.26 summarizes the key differences-in-differences regression results from the main paper under the alternative definitions. The top panel of the table summarizes the individual level results. While the magnitudes of the coefficients vary slightly across the definitions, the main finding persists that interventions reduce the probability of users to engage in hate speech. Moreover, this effect seems to be strongest for individuals who were directly targeted

Table 1.24: Intervention propensity prediction

	Logit	Lasso
Constant	-3.427***	-3.904
	(0.821)	
Share xen. act. 5 min bef. intervention	-7.769***	-1.505
	(1.933)	
Share xen. act. 100 min bef. intervention	–	-4.329
Share xen. act. 30 min bef. intervention	2.934	2.600
	(1.539)	
log(cum. act. at 100 min bef. int.)	-0.416	-0.218
	(0.264)	
log(cum. act. at 30 min bef. int.)	-0.375***	-0.382
	(0.078)	
log(act over 100-5 minutes bef. int.)	–	0.202
log(act over 30-5 minutes bef. int.)	-0.707***	-0.462
	(0.207)	
log(cum. act. at 5 min bef. int.)	–	–
log(cum. xen. act. at 5 min bef. int.)	1.612**	1.360
	(0.554)	
log(cum. xen. act. at 100 min bef. int.)	0.329	0.108
	(0.365)	
log(cum. xen. act. at 30 min bef. int.)	–	0.057
log(xen. act over 100-5 minutes bef. int.)	0.135	–
	(0.547)	
log(xen. act over 30-5 minutes bef. int.)	–	-0.147
Observations	1,081	1,081
log(λ)		-7.15
AUC	0.686	0.686

Note: The dependent variable is a dummy variable indicating whether a post from the chat log was subject to an intervention by the counterspeech group. For the LASSO regression, the regularization parameter λ is chosen in five-fold cross validation to maximize the area under the receiver operating characteristic curve (AUC).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

with a sub-level comment to their own comment by the counterspeech group. The LASSO results are again very similar to the baseline results. The bottom panel of the same table reports the results of the article-level regressions. Here, the LASSO produces a smaller and statistically insignificant effect of interventions on the number of xenophobic comments on the targeted article, explaining why I cautioned about its interpretation in the main text. The remaining results mimic the baseline.

Retaining all treatment posts

As another modification to the first step, I retain *all* treatment posts and their closest matching control post, irrespectively if they meet the other restrictions. Indeed, I cannot use all treatment posts in the baseline specification because they might not have a suitable control that is not

already matched to another treatment. Here I check that the results hold at least qualitatively when relaxing this restriction.

The descriptive statistics presented in the third supercolumn of Table 1.25 show that keeping all treatments and their closest control posts results in a larger sample in terms of both articles and users, but this comes at the cost of balance. For instance, individuals in the treatment group are significantly more active than those in the control group. Compared to the baseline sample, both subsamples are slightly less active.

This imbalance also becomes apparent in the post-level event-study plots of Figure 1.13 (blue crosses). There might be slight pretends making the common trends assumption more difficult to defend than in the baseline results. Still, the broad pattern clearly emerges that interventions were followed by a substantial increase in activity while the number of xenophobic comments and likes stayed relatively flat.

The blue crosses in Figure 1.14 show that as in the baseline specification, retaining all treatment posts yields drop in individuals' propensity to engage in hate speech in response to an intervention. The pre-intervention coefficients are slightly more volatile, however, suggesting that the treatment might be less well isolated than in in the baseline.

The differences-in-differences regression results presented in the third supercolumn of Table 1.26 confirm the key results of the baseline specification. While the magnitudes of the effects both at the individual level and at the article level are somewhat smaller than in the baseline, the qualitative patterns described in the main body of the paper remain the same.

Matching only within a time window

As a modification to the second step, I add an additional temporal restriction to the matching by first only looking for matches between articles that were considered for intervention on the same day and then by only retaining those that were discussed within two hours of the actual intervention.

Supercolumns four and five of Table 1.25 present descriptive statistics for the treatment and control groups thus obtained. Drawing control posts only from the set of posts discussed the same day as the treatment post halves the sample size and leads to slightly bigger differences between treatment and control posts even if these are not statistically significant. The user-level differences remain fairly small although the overall sample exhibits higher levels of xenophobia than in the baseline definition. Further restricting the set of possible matches to only posts that were discussed within a two hour window dramatically diminishes the sample size while simultaneously increasing the differences between treatment and control posts. Moreover, the users in both group are much more likely to write or condone xenophobic messages than in the baseline definition.

Figure 1.13 shows that the results from the article-level event-study plots presented in Section 1.5 are robust to only allowing for matches within the same day (green solid triangle) or within a two hour window (green empty triangle). Using these narrow definitions of treatment

and control group even suggests a slight decline in the number of hateful activity. The variability in the estimates of the first event-study dummy seem to stem again mostly from sample imbalance rather than actual pretrends.

Similar to the baseline, the individual-level event-study plot in Figure 1.14 indicates that users are less likely to engage in hate speech as a result of an intervention, even with temporal restrictions on the matches (triangles in the figure). The persistence of the effect might be lower than in the baseline, however. With the small sample obtained by only keeping matches within a two hour window, there is already a slight decrease in hate speech propensity in the period before the intervention which does not appear in the other specifications.

Supercolumns four and five of Table 1.26 replicate the key regression results from the main body of the paper using the additional restrictions on the matches. As with the previous robustness checks, the baseline results go through with only slight differences. At the individual level, the effect of experiencing an intervention without being targeted by a direct reply is no longer distinguishable from zero. At the article level, the effect on the total number of hateful comments on targeted articles looses its statistical significance, which is why I have cautioned about its interpretation in the main text.

Varying the number of control posts

As a second modification to the second step, I vary the number of closest matching control posts. In the baseline procedure, I keep up to three of the closest matches for each treatment posts. Here I vary this number by first keeping only the closest one and then by keeping the five closest matches.

The last two supercolumns of Table 1.25 present descriptive statistics for the samples obtained with these modifications. Retaining only the closest matches in terms of intervention propensity results naturally in a much smaller sample in terms of both posts and users. While in terms of the post-level observables, treatment and control group are even more closely comparable, the characteristics of individuals are less well aligned, which is probably explained by a 25% drop in sample size. Retaining the five closest matches, on the other hand, significantly increases the sample size. With this definition, control posts have somewhat lower levels of activity compared to treatment posts. Treatment and control users remain quite comparable.

Again, the article-level results that interventions lead to an increase in activity while the number of xenophobic comments does not change much remain robust, as can be seen from Figure 1.13. Including the five closest control posts for each treatment post leads to almost the same results as the baseline despite a much larger sample (empty yellow squares), but the common trends assumption becomes more difficult to defend using more restrictive definition (solid pink square). In general, the pattern that emerges is that the smaller the samples, the more strongly pretends appear in the graphs.

This robustness check too leads to a similar pattern in terms of the impact of interventions on individuals' future behavior (squares in Figure 1.14). There is a sharp decline in individuals

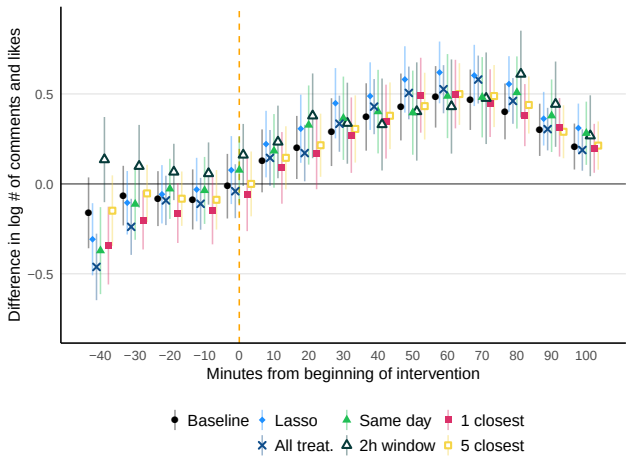
propensity to write or condone xenophobic comments following the treatment. When retaining up to five control posts for each treatment post, this effect seems to be even slightly more persistent than in the baseline.

The last columns of Table 1.26 reproduce the main regressions when retaining more or fewer control posts. The baseline results appear to be qualitatively robust for both the individual level regressions presented in the tpo panel and the article level regressions in the bottom panel. Keeping up to five control posts also quantitatively matches the baseline results, while restricting the number of controls to one per treatment lowers the magnitudes of the individual-level effects.

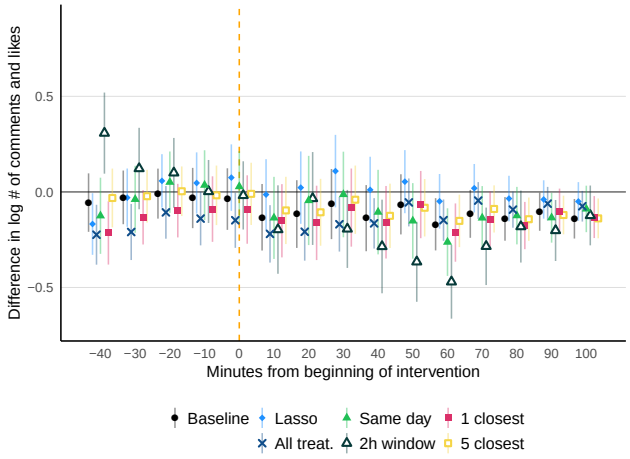
In sum, the robustness checks presented in this section leave me confident that the results of my paper are not driven by selecting specific posts to be included in the analysis but can be obtained using different selection procedures.

Figure 1.13: Robustness check on post-level event-study plots

(a) All comments and likes

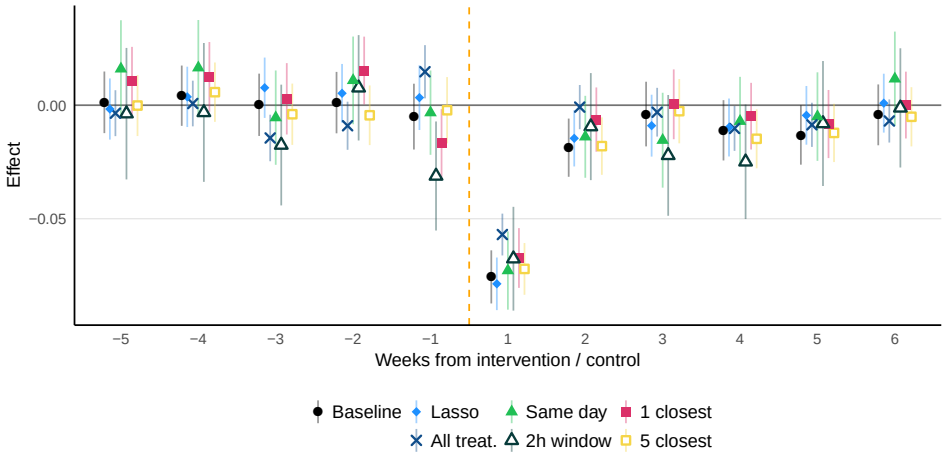


(b) Xenophobic comments and likes



Note: Event-study graph corresponding to Figure 1.9 (baseline) with alternative ways of defining treatment and control posts. "LASSO" uses the LASSO to compute treatment propensities. "All treat" keeps all treatment articles and the control articles with the closest propensity. "1 closest" and "5 closest" retain only the closest or the five closest matching potential control posts respectively. "Same day" and "2h window" only allow for matching control posts that were discussed the same day as the treatment post or within a two hour window of the treatment post respectively.

Figure 1.14: Robustness check on user-level event-study plot



Note: Event-study plot graphs corresponding to Figure 1.6 (baseline) with alternative ways of defining treatment and control users. "LASSO" uses the LASSO to compute treatment propensities. "All treat" keeps all treatment articles and the control articles with the closest propensity. "1 closest" and "5 closest" retain only the closest or the five closest matching potential control posts respectively. "Same day" and "2h window" only allow for matching control posts that were discussed the same day as the treatment post or within a two hour window of the treatment post respectively.

1.C.5 Interventions attract new users to targeted articles

In Section 1.5 I document that interventions by the counterspeech group increase the activity levels in two ways. First, there is the mechanical effect of the intervention which consists precisely of users writing couterspeech messages. Second, there is a less obvious ripple-on effect consisting of activity by users that did not directly participate in counterspeech interventions before.

Here, I disentangle the drivers of the ripple-on effect by reporting the results of two additional sets of regressions. First, I repeat the event-study analysis using the log of the number of users who comment on the specific article for the first time as a dependent variable. Panel (a) of Figure 1.15 shows that the intervention announcement leads to an influx of new users to the article, which is to be expected given that the intervention takes place. It also shows that with a slight time delay, there is an influx of users who have not previously been active in a counterspeech intervention.

Second, I plot the event-study graph for the activity levels of users who had already commented or liked a comment on the article before the intervention was announced. As can be seen in panel (b) of Figure 1.15, the intervention is not associated with an increase in activity of those users. If anything there seems to be a slight decrease in their activity levels more than one and a half hours after the start of the intervention, which would be consistent with the individual level results presented in Section 1.4.

Taken together, these two pieces of evidence suggest that the ripple-on effect is in fact driven by new users who are attracted to the article by the counterspeech intervention rather than by users who were already actively commenting on the article and try to shout back at the

Table 1.25: Sensitivity analysis: descriptive statistics

	Baseline			Lasso			All treat.			Same day			2h window			1 closest			5 closest		
	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	
Post-level comparison																					
Comments	99.2	85.8	-13.4*	103.8	88.6	-15.2*	112.8	96.7	-16.2*	90	74.5	-15.5	92.8	80.4	-12.4	98.9	88.1	-10.8	98.3	80.7	-17.6**
Reactions	399.2	395.9	-3.3	401.5	399.3	-2.2	441.7	475.2	33.5	299.4	284.8	-14.5	378.6	302.8	-75.8	424.9	424.8	0	397	368.2	-28.8
Comments & reactions	498.4	481.7	-16.6	505.3	488	-17.3	554.5	571.8	17.3	389.4	359.4	-30	471.4	383.2	-88.2	523.7	512.9	-10.8	495.3	448.9	-46.3
Users	253.9	255.6	1.7	262.5	258	-4.5	285.6	292.1	6.5	211.9	198.1	-13.8	246.8	205.9	-40.9	266.5	271.1	4.6	252.1	239.7	-12.4
Xen. comments (%)	28.1	27.6	-0.5	27.9	26.6	-1.3	20.6	21.9	1.3	24.8	23.5	-1.4	23.7	23.9	0.2	28.2	28.6	0.5	28	27.2	-0.7
Comments with tags only (%)	0.9	0.8	-0.1	0.9	0.8	0	0.8	0.9	0.1	0.9	0.7	-0.2	0.8	0.6	-0.2	0.7	0.7	0	0.9	0.9	0
Observations	178	370	548	176	379	555	312	312	624	96	170	266	53	76	129	170	170	340	176	494	670
User-level comparison																					
Avg. weekly activity	5.06	4.94	-0.12	4.91	4.77	-0.14	4.63	4.18	-0.45***	8.22	8.24	0.02	11.59	11.90	0.31	6.77	6.20	-0.57***	4.66	4.54	-0.12
Avg. weekly xen. activity	0.40	0.40	0.00	0.39	0.39	0.00	0.36	0.33	-0.03***	0.72	0.78	0.06**	1.07	1.22	0.15***	0.59	0.55	-0.04**	0.37	0.36	-0.01
Share of weeks w. activity	0.70	0.69	-0.01	0.70	0.69	-0.01	0.67	0.66	-0.01	0.77	0.77	0.00	0.79	0.79	-0.01	0.73	0.73	0.01	0.69	0.69	0.00
Share of weeks w. xen. activity	0.25	0.23	-0.02	0.25	0.23	-0.02	0.22	0.20	-0.02	0.32	0.32	0.00	0.36	0.38	0.01	0.29	0.30	0.01	0.23	0.22	-0.01
# of commented media outlets	4.30	4.23	-0.07**	4.26	4.21	-0.05	4.10	4.12	0.02	4.80	4.81	0.01	5.29	5.27	-0.03	4.60	4.52	-0.09**	4.18	4.15	-0.03
Observations	20,342	61,508	81,850	21,373	61,234	82,607	38,238	53,992	92,230	14,458	32,459	46,917	10,855	18,847	29,702	26,099	34,524	60,623	18,640	71,629	90,269

Note: Descriptive statistics for various alternative definitions of treatment and control group. The top panel of the table corresponds to Table 1.4, the bottom part to Table1.5. The first supercolumn reproduces the baseline results from the main body of the paper. In the second supercolumn, treatment propensities were computed using the LASSO. In the third supercolumn, all treatment articles were retained. The fourth supercolumns only allows treatment posts to be matched to potential control posts that were considered on the same day. The fifth supercolumn decreases this time window to two hours, i.e. only the exact alternatives that the counterspeech group considered for intervention. In the sixth supercolumn treatment and control groups have been restricted to only retain the closest match in terms of predicted intervention probability (compared to the closest three in the baseline). The last supercolumn restricts them to the closest five matches.

Table 1.26: Sensitivity analysis: regressions

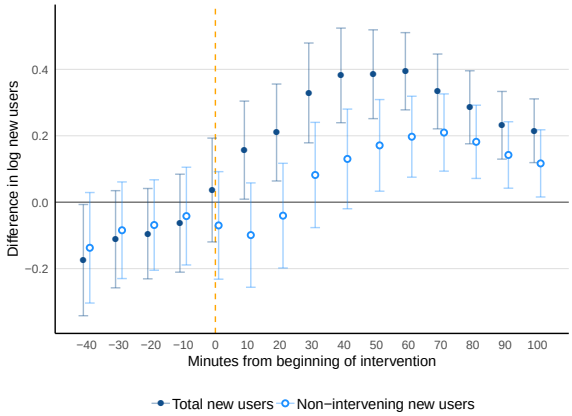
		Baseline		Lasso		All treat.		Same day		2h window		1 closest		5 closest	
Dependent var.	Coefficient	Est	Se	Est	Se	Est	Se	Est	Se	Est	Se	Est	Se	Est	Se
User-level regressions															
$\mathbb{1}\{xen.\ activity\}$	Intervention	-0.052^{***}	0.005	-0.052^{***}	0.005	-0.028^{***}	0.004	-0.044^{***}	0.006	-0.025^{***}	0.008	-0.031^{***}	0.005	-0.056^{***}	0.005
$\mathbb{1}\{xen.\ activity\}$	Intervention	-0.016^{*}	0.007	-0.015^{*}	0.006	-0.008	0.005	-0.016	0.009	-0.003	0.011	0.002	0.007	-0.018^{**}	0.007
$\mathbb{1}\{xen.\ activity\}$	Int. × SLC	-0.070^{***}	0.009	-0.071^{***}	0.008	-0.040^{***}	0.006	-0.050^{***}	0.011	-0.037^{**}	0.013	-0.060^{***}	0.008	-0.073^{***}	0.009
log(activity)	Intervention	-0.068^{***}	0.008	-0.077^{***}	0.008	-0.030^{***}	0.006	-0.066^{***}	0.010	-0.061^{***}	0.013	-0.018^{*}	0.008	-0.081^{***}	0.008
Post-level regressions															
log(xen. activity)	Intervention	-0.136^{*}	0.064	-0.074	0.067	-0.044	0.058	-0.103	0.087	-0.237	0.138	-0.089	0.078	-0.144^{*}	0.062
Share xen.	Intervention	-0.030^{**}	0.011	-0.033^{**}	0.011	-0.021^{*}	0.009	-0.043^{**}	0.014	-0.043^{**}	0.016	-0.032^{*}	0.013	-0.036^{***}	0.010
User share xen.	Intervention	-0.010^{***}	0.002	-0.010^{***}	0.002	-0.012^{***}	0.002	-0.009^{**}	0.003	-0.010^{*}	0.005	-0.010^{***}	0.002	-0.010^{***}	0.002

Note: Key regression results for various alternative definitions of treatment and control group. The rows in the top panel correspond to the user-level regression reported in Table 1.6 column 2, Table 1.7 column 2 and Table 1.8 column 3. The bottom panel corresponds to the post-level regression results contained in Table 1.11. The first supercolumn reproduces the baseline results from the main body of the paper. In the second supercolumn, treatment propensities were computed using the LASSO. In the third supercolumn, all treatment articles were retained. The fourth supercolumns only allows treatment posts to be matched to potential control posts that were considered on the same day. The fifth supercolumn decreases this time window to two hours, i.e. only the exact alternatives that the counterspeech group considered for intervention. In the sixth supercolumn treatment and control groups have been restricted to only retain the closest match in terms of predicted intervention probability (compared to the closest three in the baseline). The last supercolumn restricts them to the closest five matches. The dependent variables in the bottom two rows are computed by excluding users who already participated in counterspeech interventions previously, as explained in the main body of the paper.

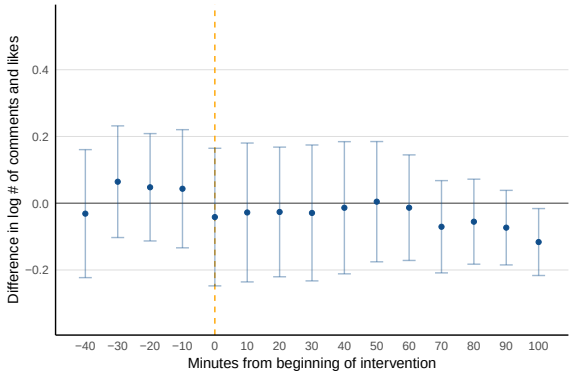
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.15: Event-study plots on drivers of ripple-on effect

(a) Number of users who are new to the article



(b) Activity by users active before intervention



Note: Event-study graph corresponding to regression (1.5) with 95% confidence intervals based on post-clustered standard errors. Regressions include time-trends and post dummies. The dependent variable in the left panel is the number of users who comment on a given post for the firs time. Non-intervention users include all users who did not write a counterspeech message in the ongoing or a previous intervention. The dependent variable in the right panel is the log number of comments and likes written by users who had already written or liked a comment on the article prior to the announcement of the intervention.

counterspeech group.

1.D Example timeline of an intervention

In order to illustrate the process that leads to a counterspeech intervention, this section provides a specific example of the first intervention of the day on August 8, 2017. In this case the article on *N24/Welt* becomes a treatment post while the article published by *Focus Online Politik* becomes a candidate for the control group.

Figure 1.16: Timeline of first intervention on Aug 8, 2017

8:05 

Focus Online Politik posts article on police intervention at refugee shelter in Bautzen

8:18 

N24 posts article on social democrat saying “Nazis more frightening than refugees to most people”

8:35 Group moderators notice article on *Focus Online*

8:38 Group moderators notice article on *N24*

9:25 

Moderators announce Intervention on article on *N24* on group’s facebook page and members start to write counterspeech comments.

1.E Content of the counterspeech comments

To disentangle what exactly drives the effect of the counterspeech interventions, it is useful to look at the messages that were sent by members of the counterspeech group during these interventions. I distinguish between top-level comments that are written in response to the media post, and sub-level comments which are written directly in response to another user's comment.

I manually classified approximately 600 comments into different content categories. The results of this exercise are reported in Figure 1.17. Each comment was assigned to at least one of the following content types:

- Arguments against hateful statements (*Argument*):
  Author provides a common-sense based argument against hateful comments or a specific hateful comment. Example: "Our ancestors were immigrants, too. So what is your problem with immigrants?"

- Disagreement with hateful statements without arguments (*Opinion*):
  Author voices her disagreement with the hateful comments without providing arguments for her view. Example: "Sad! I did not expect to live to see the day where there is so much xenophobia in this country!"

- New piece of Information to debunk hateful comment (*Info*):
  Author provides a new fact, statistic or source material to the debate. Example: "There are 500 Million people in the EU and we are talking about 40.000 refugees here. Those who claim to feel no longer at home in their own country massively distort the facts."

- Injunctive norm statement (*Injunctive*):
  Author makes a statement invoking a moral norm that ought to be followed. Example: "I do not want racism to become acceptable in this country. It is scientifically baseless, morally wrong and destructive. Racism hurts all of us"

- Descriptive norm statement (*Descriptive*):
  Author argues that the majority of people do not accept hateful comments. Example: "I disagree with you but I accept your opinion. It illustrates, however, why social networks are problematic. In surveys, only 17% of Germans are in favor of the death penalty. The likes for this comment give the wrong impression that the share is much higher."

- Call for more facts (*Debate*):
  Author calls for a more fact based debate, to read the article before commenting or to hold back with sweeping statements until more facts are available. Example: "You should back your speculative statements with facts, otherwise they are worthless!"

- Complaint about the article itself (*Complaint*):
  Author complains about sensationalism or shortcomings of the article. Example: "N24 seems to love to draw attention by publishing incomplete articles that draw attention to specific parts of the population."

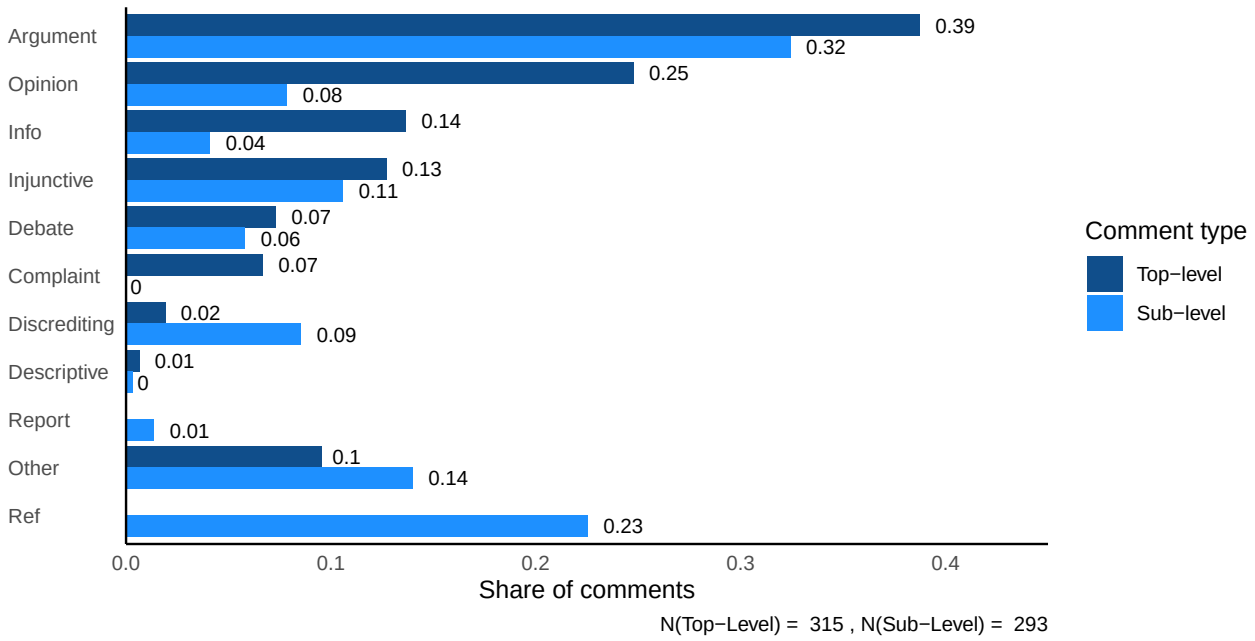
- Discredit or insult hateful user (*Discrediting*):
  Author launches ad hominem attack on users who made hateful comment or tries to discredit them. Example: "No that's not a fact, it's only a thing one of the five voices in your head is telling you."

- Threat of complaint or legal action (*Report*):
  Author tells users of hateful comments that they can or will be reported to Facebook, an employer, or law enforcement. Example: "I will report your comment."

- Unintelligible reference to another comment (*Ref*):
  Author responds or refers to another comment that makes it not clearly interpretable in the sense of these categories. Example: "That is what I am reproaching you"

- Other comments (*Other*):
  Author's comment does not fall into any of the categories above. Example: "So you are sitting there in the comfort of your own home and feel entitled to comment."

Figure 1.17: Content of counterspeech comments



Note: Manually classified comments by participants in counterspeech interventions by type of comment and content of the message.

Chapter 2

Measuring Image Concern

This chapter is joint work with Emeric Henry. It has been published as "Measuring Image Concern" in the Journal of Economic Behavior and Organization, 2019 (160).

2.1 Introduction

Individuals behave differently when their choices and actions can be observed by others. This fact is now well documented empirically (Ariely et al. (2009), Andreoni and Petrie (2004), Bursztyn and Jensen (2015)) and some important theoretical implications have been drawn (see for instance Bénabou and Tirole (2006), 2012, Ellingsen and Johannesson (2008), 2011 or Andreoni and Bernheim (2009)). Yet little is known about the drivers or the consequences of image concern. One of the main reasons for this gap in the literature is that there is currently no systematic way of measuring individual sensitivity to perceptions by strangers.[1]

The first goal of this chapter is to propose an experimental game designed to measure image concern at the individual level. A key feature of its design is its ability to identify image concern separately from other social preferences. The second goal is to use this game to examine whether image concern is linked to other social preferences.

The *image concern game* we propose involves three players: a dictator (he), a recipient and an observer (she). The dictator determines how much money to transfer to a lottery with two possible outcomes: success, in which case the recipient receives a given amount of money, or failure, in which case the recipient receives nothing. The more money the dictator transfers, the higher the chances of success. The dictator takes his decision knowing that the observer will be informed of the outcome of the lottery. Before the lottery is actually run, the dictator has to reveal his willingness to pay to remain anonymous (in an incentive compatible way), i.e. for his picture not to be revealed to the observer in case the lottery is a failure. The recipient never sees any pictures. The observer sees only the outcome of the lottery, not the amount the dictator actually transferred.

There are two main aspects that drive the structure of this game. First, image concern is easily measured by the willingness to pay to remain anonymous in case the recipient remains empty-handed. Second, if some reasonable properties of the utility function are satisfied, this measurement proves independent of other social preferences including altruism. In case the dictator does not remain anonymous, the observer does not find out how much was contributed to the lottery, only that the lottery was a failure. Thus, the inference the observer makes when she sees the picture is an updated belief on the characteristics of the dictator conditional on the fact that the lottery was a failure, and this belief cannot be conditioned on the actual amount transferred. Separating our measure from other social preferences is essential to understand the specific drivers of image concern and to show how it correlates with these other dimensions of preferences.

The game is sufficiently portable to be used in future lab or lab-in-the-field experiments to yield a measure of image concern that can be correlated with other experimental outcomes. We made sure that the game did not require complicated repeated interactions and could even be run without the different parties being present at the same time, as long as the authenticity of

[1] Heterogeneity in image concern needs to be measured to understand how it affects behavior. It is also an important element in theoretical models such as Ali and Bénabou (2016).

the participants' photos could be ensured. It is, however, less portable than other games aimed at measuring social preferences such as the trust game or the dictator game. As any setup aimed at measuring image concern will require a mechanism to vary the degree of anonymity, for instance by using pictures, this seems inevitable. We will discuss this aspect further in our conclusion.

Running this game in the lab, we find substantial heterogeneity: about one third of the participants chooses not to pay anything, while one third gives even large amounts to remain anonymous. We show that few characteristics of the observer significantly impact the willingness to pay to remain anonymous. This is encouraging evidence of the portability of the setup.[2] Nationality seems to be the exception: Non-French individuals pay significantly less for anonymity when facing other non-French observers and slightly more when observed by French observers, a fact linking nicely to the literature on discrimination.[3] One possible interpretation is that non-French participants fear that due to prejudice, French observers will interpret a failed outcome of the lottery more adversely than non-French observers.

We validate our measure of image concern in three different ways. First, we show that it significantly correlates with a survey question administered at the end of the experiment.[4] Second, we use a simple model to derive an implication of image concern and show that it stands in the data: more image concerned individuals transfer more in the first lottery to avoid situations where they would have to pay for anonymity.[5] Finally, we show that in an infinitely repeated prisoner's dilemma game, more image concerned individuals adapt their behavior to the social norm much more than others when playing in the presence of observers.

This last element of validation leverages an infinitely repeated prisoner's dilemma game, which participants played in the second phase of the experiment. In half of the sessions the game was played with observers, in the others without. We asked observers to rate the behavior of those they observe after each round so as to document what actions are judged positively by the community and identify the prevalent social norm. Using the repeated games, we first show that more image concerned individuals, when not observed, tend to cooperate less than others. We argue that this is evidence in favor of the fact that more image concerned individuals tend to be more selfish. Second, as mentioned above, comparing treatments run with observers to those without, we can show that more image concerned individuals correct their behavior in the direction of the social norm more than others – at least when they are observed by others.

[2]Indeed, if the experiment is run in different settings, different observers will be used. This evidence suggests that the measurements are not sensitive to this fact.

[3]We ask a survey question about nationality and not race, since questions on race are not allowed in France. Most non-French participants are from former French colonies in North Africa.

[4]There is unfortunately no well established question aimed at measuring image concern, contrary to the case of trust where the "Interpersonal Trust" question ("Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?") is systematically used. We thus constructed our own question: "It is important for me not to be perceived as selfish."

[5]This suggests that image concerned individuals avoid situations where they risk being exposed. This is coherent with the results of Dellavigna et al. (2012), who show that when the date of a door-to-door fundraising visit is announced, people try to avoid being present. Our individual measurement of image concern allows us to show direct evidence of such a mechanism whereby image concerned people avoid situations where their image is at risk.

In the last section of the chapter we discuss different properties of the image concern game. We also show that our main results are robust to adjusting for multiple hypothesis testing, following an approach proposed by Benjamini and Hochberg (1995).

Our research is closely connected to the empirical literature on the influence of being observed, using both field and lab experiments. We differ in both our goal and approach. The goal of most of these papers is to document the average influence of being observed by comparing average differences between treatments. We, on the other hand, are interested in individual measurements and individual consequences. Ariely et al. (2009), for instance, compare effort levels in treatments that varied in three dimensions: Subjects were either observed or unobserved, received monetary incentives or not and contributed either to a "good cause" (Red Cross) or a "bad one" (NRA). They find that being observed increased effort levels only when subjects did not receive monetary incentives and only when they volunteered for a good cause. Andreoni and Bernheim (2009) derive a model that can explain the strong prevalence of 50-50 splits in the dictator game by individuals' desire to be perceived as fair by others. They show that their model is confirmed by data from a modified version of the dictator game in which transfers to the recipient are sometimes determined randomly, rather than by the dictator. Ekström (2012) finds that norm compliance can be increased even by subtle clues of observation, such as pictures of human eyes. Riyanto and Zhang (2015) deviate from this tradition by eliciting the willingness-to-pay of participants in a dictator game to avoid seeing other people's opinion on their action. However, their design can only capture the dictator's self-image concern and not their concern for others' opinion of them.

There is also a strand of the literature documenting consequences of being image concerned. Lacetera and Macis (2010) show that blood donors increased the frequency of donations in order to reach thresholds which would earn them prizes, but only if the prizes were publicly announced and awarded. This suggest that image concern can be an important driver of unselfish actions. Dellavigna et al. (2012) show that notifying residents in advance of a door-to-door fundraiser significantly decreases the share of households opening doors, one possible interpretation being that image concerned individuals attempt to avoid the pressure. Bursztyn and Jensen (2015) show that image concern can affect educational choices and show the importance of the observer's identity.

Further evidence on the consequences of image concern has been gathered using laboratory experiments, where typically behavior is compared across treatments with and without observers. Andreoni and Petrie (2004) find that contributions in a public goods game increased when the players were not anonymous. Dana et al. (2006) offer participants a costly possibility to opt out of a dictator game and show that giving in the dictator game is in part motivated by participants not wanting to appear selfish. In the same spirit, other contributions find that providing options for the participants to overcome their moral dilemmas significantly lowers transfers (Rege and Telle (2004), Samek and Sheremeta (2014), Dana et al. (2007)).

We point out one branch of the literature that tries to find individual proxies for image concern. Carpenter and Myers (2010) use data on the purchase of vanity plates by firefighters

that make them identifiable as such at all times. They show that this proxy can predict higher responses to emergency calls but has no effect on less visible activities such as training. In a study of Wikipedia, Algan et al. (2013) use the size of the contributor's page and the extent to which they choose to display awards as a proxy for image concern.[6] We share the goal of these papers to find individual proxies for image concern but try to determine a less context specific measure exploitable in a wide variety of settings.

Our approach is similar to some extent to the approach in the literature on trust. Analogously to Glaeser et al. (2000), Fehr et al. (2003) and Sapienza et al. (2013), we compare measurements of preferences obtained by survey questions to those resulting from laboratory experiments. Our results also link us to the literature on racial discrimination and bring a new twist by documenting that non-French subjects are only image concerned when facing French individuals. There is a growing literature experimentally studying issues of discrimination and prejudice (for instance Fershtman and Gneezy (2001)). Here what seems to play a role is the fear of prejudiced reactions.

Finally, the second phase of the experiment relates to the literature on infinitely repeated games in the lab (Dal Bó and Fréchette (2011) and Dal Bó (2005) among others). To the best of our knowledge, it is the first time an infinitely repeated prisoner's dilemma is played with observers, a side contribution of our paper.[7] In addition to our analysis of image concern, our study also allows for a better understanding of the social norms governing those games, using the ratings by observers of the behavior of participants.

The remainder of this article is organized as follows. Section 2 introduces the image concern game and a conceptual model to analyze it. Section 3 presents its results. Section 4 analyzes the relation between our measure of image concern and behavior in an infinitely repeated prisoners dilemma game. Section 5 discusses robustness and potential concerns and section 6 concludes.

2.2 Measuring image concern: concept and procedure

Although being observed by strangers has been shown to impact behavior and many models include a term corresponding to image in the utility function, there is no unified concept, model or even terminology. We therefore start in section 2.2.1 by defining the concept we want to measure, then introduce in section 2.2.2 the image concern game we propose to perform this measurement and analyze the game in section 2.2.3. Section 2.2.4 details the setup of the experiment.

[6] Algan et al. (2014) in an analysis of open source software programmers use the answer to a survey question to identify image concern.

[7] Charness et al. (2007) do examine a prisoner's dilemma with observers, but without repeat interaction and where the observers are group members who have a stake in the game. Other papers have focused mainly on punishment by third-parties for norm enforcement, rather than observation alone (see for instance Fehr and Fischbacher (2004)). In our setup observers have no possibility to punish players. Sutter et al. (2009) use a much weaker form of observation, where observers only know about decisions and payoffs but do not see players' pictures or anything that could identify them.

2.2.1 Conceptual framework

We define the image concern of an individual as the degree to which anonymous strangers' opinion of him affects his utility. Specifically, the utility of individual i, when he consumes c_i and others consume c_{-i} and when the image others have of him is denoted R_i (defined formally below), is given by:

$$U_i = V_i(c_i, c_{-i}) + \mu R_i \tag{2.1}$$

The term μ measures image concern, the dimension of preferences we want to measure. This has been sometimes called "concern for social image" (Andreoni and Bernheim (2009)) or "image motivation" (Ariely et al. (2009)). Note that the utility function can include other social preferences such as altruism, since the consumption of others directly enters the utility function.

Suppose the characteristics or type of the individual is multidimensional $v^i = (v_1^i, .., v_K^i)$. The characteristics could be altruism, reciprocity or other individual characteristics that could potentially influence the shape of the function V_i. The characteristic k for individual i is drawn from the distribution Φ_k^i. The image term R_i corresponds to the beliefs others hold on i's characteristics. Specifically, we assume that for a given characteristic k, R_i is the difference between the expected value of characteristic k for individual i and the average value of the characteristic in the population v_k^0. Finally, we assume that individual i might not care in the same way about the image he conveys on the different characteristics. We therefore introduce weights γ_k^i on characteristic k, so that R_i is given by:

$$R_i = \sum_{k=1}^{K} \gamma_k^i \left(E\left[v_k^i\right] - v_k^0 \right) \tag{2.2}$$

The expectation $E\left[v_k^i\right]$ is affected by observable actions taken by individual i as in the case of the image concern game we introduce in the next section, where the observer updates her beliefs about the sender based on the outcome of the lottery if she observes it.

2.2.2 The image concern game

The image concern game we propose is played between three players: the dictator (he), the recipient and the observer (she). The game is played as follows:

1. The dictator sees the photo of the observer but the observer does not see any pictures yet. The recipient never sees any pictures (neither of the observer nor of the dictator) throughout the game.

2. The dictator receives 100 tokens. He decides how much to allocate to a lottery. The lottery has two possible outcomes: success, in which case the recipient receives 50 tokens or failure, in which case the recipient receives nothing. For each token paid by the

dictator, the chances of success increase by one percent, i.e. if the dictator gave an amount $X \in (0, 100)$, the probability that the lottery is a success is $X/100$.

3. Before the lottery is run, the dictator chooses the maximum amount b he is willing to pay to remain anonymous in case the lottery results in a failure. To ensure truthful answers, we use a Becker-DeGroot-Marschak (BDM) type mechanism.[8]

4. The lottery is carried out:

   (a) If it is a success, the recipient receives 50 tokens and the picture of the dictator appears on the screen of the observer.

   (b) If it is a failure, the recipient receives nothing and the BDM mechanism comes into play. A random number $r \sim U(0, 100)$ is drawn. If $r \leq b$, the dictator pays r and remains anonymous (the observer does not see the dictator's picture). If $r > b$, the dictator pays nothing and the observer sees the picture of the dictator. In both cases, the observer learns that the lottery outcome was a failure.

   No matter the result of the lottery, neither the observer nor the recipient ever learn the amount actually chosen by the dictator in either step. They are only informed about the outcome of the lottery.

As suggested in the introduction, there are several key ideas that underly the setup of this game.

First, the individual image concern can be measured by the willingness to pay b to remain anonymous, chosen in step 3. This is formally shown in Proposition 1 for the case of separable cost functions. Participants who are image concerned would prefer to remain anonymous, since if the recipient did not receive any money in the lottery, the observer would infer that they behaved selfishly.

In practice, image concern can also lead individuals to pay to make themselves visible when they have behaved generously. For example, Carpenter and Myers (2010) use the purchase of vanity license plates as an indicator of sensitivity to image concerns. However, we deliberately set up the experiment using avoidance of negative perception rather than bidding for positive image in order to rule out that bidding itself could be adversely interpreted. We were concerned that participants would not want to be perceived as showing off if they paid to be visible, blurring the measurement of image concern. We thus opted for a set up that corresponds to situations in which people incur cost of effort in order to cover up behavior that might be interpreted negatively by others.[9]

[8] This method of incentive compatible WTP elicitation was introduced by Becker et al. (1964) and is very common in the literature. To avoid concerns that it is not well understood by participants, the instructions clearly stated that the best strategy is to honestly report WTP.

[9] It would be a fruitful avenue for future research to test whether the alternative experimental setup, letting participants pay to be visible, would yield different results.

Second, we chose to have three players, rather than just a dictator and a recipient, in order to withhold the identities of dictators and recipients from each other. We believe we needed the dictator to see a picture to personify the potential observer. Without a picture, the fact of being observed would have been too abstract. Given this need for a picture, we preferred adding a third party as observer for three reasons. First, if the dictator's picture was shown directly to the recipient, dictators might worry about retaliation outside the lab. Retaliation by a third party not directly affected by the dictator's decision seems far less likely. Second, this setup guarantees that dictators are not influenced by the recipients' characteristics visible on the picture (such as the perceived needs). Finally, it allowed us to test for the effect of observer characteristics, such as gender. Note that if these issues do not appear to be of first order for the experimenter, the experiment can be run in a simpler version with just a dictator and a recipient. This may be the case in experimental settings where retaliation outside the lab is unlikely, for instance in large online surveys. As we report in section 2.5, observer characteristics seem to play a minor role.

Third, the decision to pay for anonymity is separated from the amount actually transferred in the lottery by the dictator. Regardless of how much the dictator gave in step 1, the inference that an observer makes about the dictator's generosity when she sees a failure is the same since she does not observe the actual amount transferred. This is also clarified below in Proposition 1. Intuitively, the fact that the measurement is not confounded by altruism becomes clear when considering two dictators with the same image concern but different levels of generosity. In our setting, the two dictators will give different amounts in the lottery but will bid the same way for anonymity. We could have chosen a setting where dictator and recipient play a classical dictator game and the dictator first has to bid for anonymity given that the amount transferred will be revealed to the observer. In such a case, the more generous dictator would still transfer more than the other in the lottery, but would then bid less for anonymity since he would have less to be ashamed of. We would thus mistakenly conclude that the first dictator was less image concerned. Our game, at a slight cost of complexity, is designed to overcome this potential issue.

Fourth, image cannot lead to future material payoffs since the participants are randomly rematched in later stages in the game. Payoffs from future interactions outside the lab are unlikely as most of the lab participants do not know each other and we control for this factor when they do.[10]

2.2.3 Analysis of the game

We clarify the claims made above by deriving theoretically the equilibrium choices in the experiment if the utility of participants is given by equation (2.1).

The dictator has two choices to make: the amount X he transfers to the lottery and the

[10]In the literature, for instance Bénabou and Tirole (2006), the reputation term is allowed to cover all the different dimensions, image concern, self image or reputation payoff.

amount b he bids to remain anonymous (both variables are normalized by 100, so that for instance the probability of winning the lottery is X). We put more structure on the function V_i introduced in equation 2.1. We denote $v^i(1)$ the utility of dictator i net of costs and reputation if the lottery is a success (and the recipient receives the 50 tokens), while $v^i(0)$ is the corresponding value in case of a failure. Thus $v^i = v^i(1) - v^i(0)$ measures the altruism of individual i. This will be the only characteristic defining the type (i.e. the number of characteristics is $k = 1$).

We assume in the following that the cost functions of giving to the lottery and bidding for anonymity are separable. We denote $c_1(X)$ the strictly increasing and convex cost function of giving to the lottery and $c_2(b)$ the strictly increasing and convex cost function of bidding for anonymity. Individual i thus chooses X and b to maximize

$$X \underbrace{\left[v^i(1) + \mu R_1^*\right]}_{\text{payoff when lottery succeeds}} + (1 - X) \underbrace{\left[v^i(0) + (1 - b)\mu R_0^* - bc_2(b)\right]}_{\text{payoff when lottery fails}} - c_1(X)$$

where $R_l^* = E[v|l] - v_0$, where v_0 is the average altruism under the ex ante distribution, and $E[v|l]$ is the expectation of v conditional on the outcome $l \in \{0, 1\}$ of the lottery. A successful outcome brings a positive image since it signals that the dictator has likely transferred more money, $R_1^* > 0$, while a failure is a bad signal $R_0^* < 0$, as we show in the proof of Proposition 1. Note that compared with equation 2.2, we implicitly suppose that $\gamma_1^i = 1$. In any case in the data, μ and γ_1^i cannot be identified separately, since there is a single characteristic $k = 1$.

We obtain the following results:

Proposition 1: In an interior equilibrium, if the cost functions are separable, the bid for anonymity b^* of an individual is strictly increasing in image concern μ. Furthermore:

1. the bid for anonymity b^* of an individual is independent of the altruism v,
2. the transfer to the lottery X^* is increasing in the altruism v and in image concern μ.

Proof: see appendix 2.A.1

Under the assumption that the cost functions c_1 and c_2 are separable the bid for anonymity is increasing in image concern. In appendix 2.A.2 we discuss the conditions under which this results hold in the non-separable case as well. Two additional results are obtained under the separability assumption. First, our measurement of image concern is independent of altruism v. The idea behind this result is that the bid matters only in the case the lottery is a failure, and conditional on a failure, the amount X actually invested no longer plays a role. This is a key feature that drove the design of our experiment. Second, under the assumption that v and μ are not correlated, more image concerned individuals will tend to invest more in the lottery to avoid having to pay for anonymity. Indeed, more image concerned individuals know that a failed outcome will be more costly for them as they will have to pay more for anonymity. This

is consistent with the interpretation given in Dellavigna et al. (2012), who show that notifying residents in advance when a door-to-door fundraiser takes place significantly decreases the share of households opening doors. One possible interpretation for this finding is that image concerned individuals attempt to avoid the pressure. Our individual level measure of image concern allows us to show even more precise evidence of such a mechanism. This last prediction of Proposition 1 is used in section 2.3.2 as one of three methods to validate our measure of image concern. Note that it does rely on the assumption that the level of image concern is uncorrelated with altruism. A sufficiently large negative correlation could reverse the result.

2.2.4 Experimental setup

Organization of the sessions

The experiment was computer-based and there was no communication between subjects. All participants were seated in the same room, separated by screens, and briefed together. Before the experiment started, a picture was taken of each participant and fed into the experimental software, so that subject anonymity could be removed in a controlled manner. The participants were informed that the photo would be destroyed immediately after the end of the session. Each session was organized in three phases:

1. Participants played four successive and independent rounds of the *image concern game*. They were randomly assigned to be either a dictator, a recipient or an observer and informed of their assignment. They kept this role for the four rounds. At the beginning of each round, a photo of the observer was shown to the dictator (in the right panel of his screen as shown in Figure 2.6 in the appendix). At the end of each round, dictators were rematched with different observers and recipients. In each round an observer was assigned two dictators. No dictator encountered the same observer or recipient twice.

   The players were informed that they would play four rounds but that only one of them would be selected at random to determine their payoffs. Nevertheless, at the end of each round, they observed the outcome. The payoff of the dictator and the recipient depended on the dictator's choice and the outcome of the lotteries, as described in section 2.2.2. The observers, on the other hand, received a fixed payment of 40 tokens per round, independent of other players' actions.

2. Subjects played a repeated prisoner's dilemma game described below. In half of the sessions, the prisoner's dilemma games were run with third party observers, in the others without. The main goal of this additional stage is to provide an additional method to validate our measure of image concerns.

3. A survey was conducted containing the main question we use as validation for our image concern game, as well as questions on socioeconomic information. We also included a general question on risk-taking that has been shown to be strongly correlated with

incentive compatible measures of risk-preferences and to predict risky behavior (Dohmen et al. (2011)).

At the beginning of each phase, participants received a copy of the instructions, which were then read out loud by the experimenters. Participants filled out a brief questionnaire to check their own understanding and could ask questions in private. The experimenter then read out the correct answers to the questionnaire, making us confident that subjects accurately understood the instructions.

Infinitely repeated prisoner's dilemma

Players were organized in pairs and played the following prisoner's dilemma with payoffs presented in Table 2.1.

Table 2.1: Payoffs of prisoner's dilemma

	C	D
C	8, 8	0, 10
D	10, 0	4, 4

The infinitely repeated game was implemented using a random continuation rule where at the end of each round there was a probability of 7/8 that another round was played in the game.[11] After the last round of a game had ended, the participants were rematched so that no group of subjects encountered each other more than once. The participants were informed that they would play exactly three games. In practice, as in Peysakhovich and Rand (2015) and Fudenberg et al. (2012), we did not randomize the number of rounds within a game *during* the session but *before* since we wanted to compare behavior across treatments and thus wanted games of identical length. We followed the randomization chosen by Peysakhovich and Rand (2015), who also used a continuation probability of 7/8, and we chose exactly the same length as in the first three games described in their paper. Given this approach, each participant played three games, the first with 12 rounds, the second with one round and the third with three rounds.[12]

In half of the sessions, the actions of both players in the prisoner's dilemma were visible to an observer who did not have any stakes in the game. Each observer was assigned to two pairs of players. The observers saw the players' photos and computer names, as well as the decisions they made in the game. A picture of their observer and his or her computer name was visible on the players' screens while they took their decisions. Observers had to indicate whether they had met the other participants before and were asked in each round, after having observed the choice of the players, how they rated their behavior in the game.

[11] With this continuation probability, cooperation can be sustained in a subgame perfect and risk dominant equilibrium.

[12] In the instructions it was not made clear whether the randomization at the end of each round was done on the spot or had been done before (as was in fact the case). It was only stated that at the end of each round, there were 7 chances out of 8 to have another round in the game.

The payoff in this phase of the experiment was the sum of payoffs in all rounds. The observers received a flat payment of 5 tokens per round that was independent of the players' actions.

The sample

The experiment was conducted in May and September 2014 at the Laboratoire d'Economie Experimentale de Paris. The lab has access to a subject pool that comprises individuals not affiliated to any university as well as students and staff. Table 2.2 provides descriptive statistics both for the full sample and for the sample of participants for whom we have an image concern measure, i.e. who played the image concern game as dictators. The sample is fairly balanced in terms of gender and marital status and is not exclusively composed of students (61 percent of students in the dictators sample). The majority of participants is French and most of those who report not being French are from North African countries.

Table 2.2: Descriptive statistics for the sample of participants

Variable	All Participants *Share/Mean*	Dictators *Share/Mean*
Demographics		
Female	.56	.59
In a relation	.49	.58
Age	28.5	30.20
French national	.87	.85
Attitudes to Risk	5.75	5.76
Professional status		
Student	.66	.61
Employed	.21	.22
Unemployed	.09	.11
Retired	.04	.06
Highest degree achieved		
High school or less	.29	.31
College diploma	.44	.45
Master's degree	.24	.22
PhD	.04	.02
Field of study		
Economics and finance	.28	.29
Other social sciences	.15	.16
Law	.15	.11
Picture controls		
Smiles on photo	.07	.09
Frowns on photo	.02	.02
Photo blurry or eyes shut	.03	.03
Observations	260	104

Note: Descriptive statistics for the sample of lab participants. The left column reports statistics for the entire sample, the right column for those participants who were assigned to the role of dictator in the image concern game.

While the experiment was ongoing, we documented whether participants smiled or frowned

on their pictures, and whether the picture was blurry or participants had their eyes closed.[13] To accommodate privacy concerns, we deleted all pictures immediately after each session.

Each of the 13 sessions involved exactly 20 participants. In the image concern game, eight of those participants were assigned to the role of dictator, eight to the role of recipient and four to the role of observer (observers were in charge of two dictators in each round). For the prisoner's dilemma, half of the subjects played with observers and half without. Overall this gives 260 subjects, out of which 104 played the image concern game in the position of dictator. On average, subjects received €16.74 for participating in the experiment, including a fixed €4 show-up fee.[14]

2.3 Measuring image concern: the results

2.3.1 Heterogeneity in image concern

The image concern game is designed to measure image concern in a straightforward way using the willingness to pay for anonymity. In principle, only one round of the game is needed to obtain a measurement. However, we ran four successive rounds with rematching in order to assess the impact of observers' characteristics on the measure. For each individual, we thus have four individual measures that we could potentially combine in different ways. However, we find little variability in aggregate bidding behavior: The average value of μ does not vary much across rounds, except for the last round where it is slightly but not significantly lower. Individuals' bids are also highly persistent across rounds. For the remainder of the chapter, we will therefore use the bid for anonymity in the first round as an individual's measure for image concern.[15] To facilitate interpretation we divide it by its standard deviation wherever it is used as an explanatory variable.

The results of the experiment reveal significant heterogeneity in terms of image concern. The distribution of the bids for anonymity chosen in the first round is given in Figure 2.1 (a).[16] 40 percent of the sample is completely insensitive to image, i.e. does not pay to stay anonymous. On the contrary, more than 22 percent seem quite sensitive and give more than 20 tokens to stay anonymous in the case of an adverse lottery outcome. Since this is the first study to measure individual sensitivity to image using an experimental game, it is difficult to compare the distribution to existing results. As a reference point, Carpenter and Myers (2010) find that 23 percent of firefighters in their sample purchase a vanity plate for their car, which is the proxy the authors interpret to identify image concerned individuals.

Dictators' transfers to the lottery exhibit substantial heterogeneity as well (see Figure 2.1(b)). Only 20 percent of dictators do not contribute anything to the lottery that benefits

[13] The experimenter was unable to see any of the decisions participants made while he was coding these variables.

[14] The conversion rate of tokens to euros was 10 to 1.

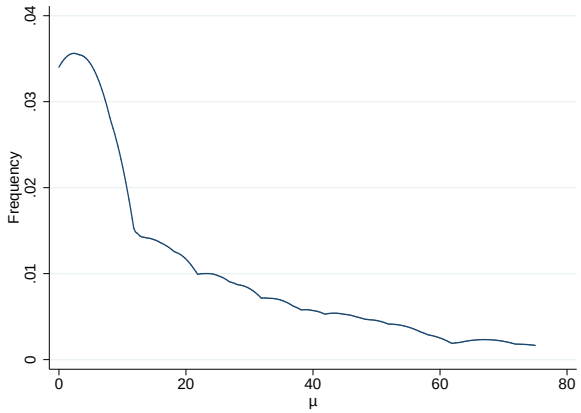
[15] Our results are mostly robust to using alternative measures such as the average bids.

[16] Its shape is very similar for the average value of the bids (see 2.7 in the appendix).

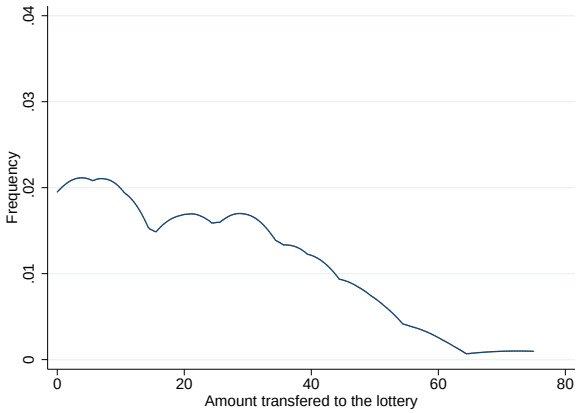
the recipient, the average amount transferred is around 19 tokens and its standard deviation is 19. While the average transfer remains largely stable across the four rounds, the share of non-contributions seems to be weakly significantly higher in the last round.[17]

Figure 2.1: Distribution of transfers and bids in the image concern game

(a) Distribution of image concern



(b) Distribution of transfers to the lottery



Note: This figure plots a kernel density estimates of (a) dictators' image concern as measured by their first bid for anonymity (Epanechnikov, bandwidth = 5.3) and (b) amounts transfered to the lottery by dictators in the first round (Epanechnikov, bandwidth = 6.4)

In our small laboratory sample, sensitivity to image concern does not seem to correlate with socioeconomic characteristics as shown in Table 2.3. In column 1 we consider the full sample, in column 2 we restrict the sample to those having bid a positive amount for anonymity, while in column 3, we use an indicator variable of whether the bid for anonymity is positive as dependent variable. The only fact that seems to emerge is that women might be more prone to bid *something* to remain anonymous, even if the amount is low. Of course, one limitation of our study is sample size. Larger samples could potentially uncover correlations that we are not able to find. An interesting direction for further study in this context would be to correlate our measure of image concern with well-established questionnaires from the psychology literature on personality traits, such as the "Big Five" (Costa and McCrae (2008)).

For most of the analyses we exclude three outliers who bid more than 90 for anonymity whereas the highest bid among the rest of the population is 75. The results are robust to the inclusion of these individuals. This leaves us with 101 individuals for whom we have a measure of image concern.

2.3.2 Validation of the game

One key question immediately arises: Are we indeed measuring image concern? There is not, as in the case of trust, a widely accepted survey question convincingly capturing the degree of image concern. We therefore constructed a question that reflects this construct: "It is important for me not to be perceived as selfish" on a 0-5 scale. We show in Figure 2.2 the average image

[17] At the individual level, transfers are highly persistent and seem not to be correlated with our survey measure of risk aversion.

Table 2.3: Explaining image concern

	Image concern *OLS*	Image concern *OLS*	Non-zero Image con. *Probit*
Female	-0.45	-6.97	0.43**
	(3.51)	(6.18)	(0.16)
Age	0.12	0.00	0.01
	(0.25)	(0.23)	(0.02)
In a relationship	2.22	7.59	-0.29
	(3.49)	(4.94)	(0.28)
Student	-8.01	-7.51	-0.65
	(7.63)	(9.56)	(0.44)
French	-3.59	-3.15	-0.71
	(6.91)	(5.62)	(0.46)
Knows the observer	5.39	3.19	0.01
	(5.88)	(8.87)	(0.41)
Constant	17.33	30.53^{+}	1.10
	(13.79)	(15.28)	(1.02)
Observations	101	61	101
Pseudo R^2	0.091	0.130	0.110

Note: Column 1 presents a regression using the bid for anonymity the first time the image concern game is played as dependent variable and socioeconomic characteristics for the full sample as explanatory variables. Column 2 restricts the sample to individuals who bid a positive amount. Column 3 is a probit regression of the indicator variable taking value 1 if the individual bid a positive amount for anonymity. The sample contains all individuals who played as dictators in the image concern game, excluding 3 outliers. Session-clustered standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

concern measure for each of the answers to this question. The more people agree with the statement, the more they were willing to pay for their anonymity in case the lottery outcome in the experiment reflected badly on them. This suggests that our measure does capture the sensitivity to the perception by others.

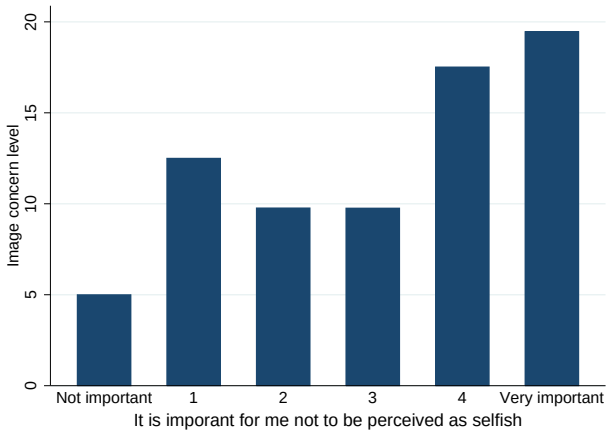
We confirm these graphical results in Table 2.4. In column 1 we present the results of an ordered logit where the dependent variable is the answer to the survey question. There is a positive and significant correlation between the bid for anonymity and the answers.[18] It is important to note that it is the only variable that can explain variations in the answer to that question. In particular, none of our socioeconomic variables turns out to significantly impact the answers.[19]

The association with the natural question in our questionnaire offers a strong initial validation of our measure of image concern. Our second method to validate the measure is to test result 2 of Proposition 1 that stated that more image concerned individuals should bid more in the initial lottery phase of the experiment to avoid situations where they will have to

[18]It is unlikely that participants answered the question in this way purely to appear consistent with their behavior in the game, since between the survey and the image concern game, other games had been played.

[19]The survey question also correlates positively and significantly with contributions to the lottery ($r_s = 0.16, p < 0.01$)

Figure 2.2: Image concern levels by answers to the statement "It is important for me not to be perceived as selfish" (scale 0-5)



Note: This figure reports the average image concern levels as measured by the first bid for anonymity. The survey question was "It is important for me not to be perceived as selfish" and participants could answer on a 0-5 scale ranging from "not important" to "very important."

pay to preserve their anonymity. We find in the second column of Table 2.4 that more image concerned individuals transfer significantly higher amounts to the lottery: A standard deviation increase in the image concern measure is associated with an increase in transfers by eight tokens. An alternative interpretation of this finding could be that image concerned individuals are just more generous. We in fact show in the next section that the correlation goes in the other direction.

Our third and final method to validate the measure is presented in section 2.4.4 where we compare sessions where the repeated game was played with observers to those where it was not. We show that more image concerned individuals, as measured by our game, react more to being observed and in particular are more likely to choose the action that observers judged positively.

2.4 Impact of image concern in an infinitely repeated prisoner's dilemma

We now turn to the analysis of the behavior of participants in phase two of the experiment, where they played the prisoner's dilemma. We will exploit both the differences across sessions (sessions with and without observers), as well as individual heterogeneity in image concern within sessions.

The purpose of this analysis is threefold. First, to establish what subjects see as the behavioral norms that prevail in the repeated prisoner's dilemma. Second, to determine whether image concerned individuals behave differently than others with respect to these social norms in the absence of observers, for instance in terms of cooperation rates. Finally, we examine whether image concerned individuals react more than others to the fact of being observed.

Table 2.4: Validation of our image concern measure

	Survey question *Ordered logit*	Transfer to lottery *OLS*
Image concern (μ)	0.50**	7.98**
	(0.18)	(2.32)
Knows the observer	-0.04	-2.98
	(0.47)	(4.71)
Economist	-0.93*	-0.98
	(0.43)	(3.64)
Response time	-0.02*	-0.00
	(0.01)	(0.06)
Female	-0.11	3.53
	(0.32)	(3.61)
Age	0.01	0.10
	(0.02)	(0.16)
In a relationship	-0.41	2.31
	(0.44)	(2.78)
Student	-0.32	-0.14
	(0.54)	(5.68)
French	-0.97	-1.79
	(0.70)	(5.74)
Observations	101	101
Pseudo R^2	0.065	0.244

Note: This table reports in the first column, the estimations of an ordered logit regression of the answers to the question "It is important for me not to be perceived as selfish" (0-5 scale), and in the second, a regression of the amount transferred by the dictator in the image concern game, regressed on image concern measures and socioeconomic characteristics. The image concern measure has been divided by its standard deviation for easier interpretation of the magnitudes. The sample contains all individuals who played as dictators in the image concern game, excluding 3 outliers. Session-clustered standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

This will provide further validation of our measure of image concern and inform us on the average effect of being observed in an infinitely repeated prisoner's dilemma game.[20]

2.4.1 Prevailing social norms in the repeated prisoner's dilemma

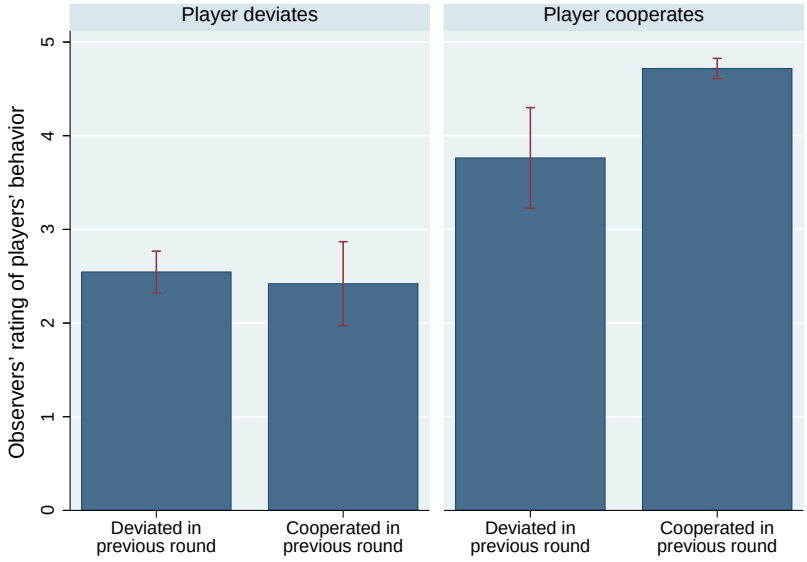
We hypothesize that more image concerned individuals adhere more closely to social norms when they are observed compared to when they are not. Showing that this is the case requires as a first step to determine what actions in the repeated prisoner's dilemma are considered appropriate by the community, i.e. what is the social norm. We are able to address this interesting and novel question by studying observers' ratings of players' actions. After each round of the game, observers were asked to rate the actions of both players. This allows us to document two social norms:

[20]Measures of image concern obtained from the survey questions perform less well at explaining behavior in this game compared to the measure obtained by the experiment and presented in this section.

- Cooperation is judged favorably by observers – independently of partners' previous actions
- Consistent, reliable cooperation garners additional approval by observers

The first point unambiguously appears in Figure 2.3: Cooperating is highly rated by the observers. However, the perception of observers is also based on a more subtle reaction to the history of play: The positive rating of cooperation is particularly strong when the player also cooperated in the previous round. The observers rate consistency in cooperation very highly. One interpretation could be that they value unconditional cooperators, who consistently avoid deviating.

Figure 2.3: Rating of behavior by observer depending on past choices



Note: This figure reports average ratings of players' decisions in the infinitely repeated prisoner's dilemma game by outside observers. 95% confidence intervals in red.

The regression analysis presented in Table 2.5 confirms these findings. We first note that a certain number of facts not linked to behavior affect the ratings. Students and those who smile in their picture are better rated than others. French observers tend to give significantly lower ratings to their fellow citizens and higher ratings to non-French participants. A possible interpretation for this behavior is that they fear being perceived as prejudiced.

In terms of observed behavior, cooperation indeed significantly increases the rating (Table 2.5). We also confirm that cooperation following cooperation in the previous round has an additional positive effect on ratings – stability is valued (column 3). On the other hand, there is no significant dependency of the rating on the partner's action in the previous round (column 2).[21] These social norms will be used in the next sections to determine how image concerned individuals react to observation.

[21]It would be natural to think that ratings would also depend on what the other player did in the past. However, both Table 2.5 and Figure 2.8 in the appendix show that there is no extra negative rating coming from a deviation that follows cooperation by the partner, i.e. there does not seem to be a judgment on betrayal of the partner.

Table 2.5: Observers' ratings in the infinitely repeated prisoner's dilemma game

	Rating		
Player cooperates	2.94**	2.47**	1.78**
	(0.54)	(0.66)	(0.66)
Player French	1.11	1.06	1.19
	(0.81)	(0.73)	(0.74)
Observer French	0.66	0.53	0.62
	(0.51)	(0.53)	(0.50)
Observer French × player French	-1.40+	-1.28+	-1.40+
	(0.80)	(0.73)	(0.73)
Player is student	1.21+	1.24+	1.22+
	(0.66)	(0.65)	(0.63)
Player smiles	2.37**	2.16**	2.00*
	(0.84)	(0.82)	(0.80)
Deviate after partner cooperated		0.29	
		(0.40)	
Cooperate after partner cooperated		0.75	
		(0.56)	
Deviate after player cooperated			-0.15
			(0.35)
Cooperate after player cooperated			1.60*
			(0.69)
Observations	520	520	520
Pseudo R^2	0.242	0.246	0.253

Note: This table reports the estimations of an ordered logit regression of the ratings given by the observer (1-5 rating) on the player's characteristics and behaviors. We control for players' and observers' age and gender, for whether player and observers knew each other prior to the experiment, as well as for attitudes towards risk. As a robustness check, Table 2.10 reports the results of a probit model predicting whether an observer gave the highest rating. The results are very similar. Standard errors in parentheses are clustered at the individual level. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

2.4.2 Cooperation rates of image concerned individuals when not observed

In environments where they are not observed, more image concerned individuals are more likely to defect – they fail to comply with the social norm. To show this, we exploit the variation within sessions of the image concern parameter to determine whether more image concerned individuals behave differently than others, particularly in terms of cooperation rates.

In Table 2.6 we report the results of a probit regression of the cooperation variable explained by socioeconomic characteristics and by the image concern parameter, restricting attention to treatments where the prisoner's dilemma was played without observers. Regardless of whether we focus on all rounds (column 1), only the first round of each game (column 2) or exclusively on the last game (column 3), we find that the more image concerned individuals cooperate significantly less. This is particularly striking since no socioeconomic characteristic has consistently significant explanatory power.

Table 2.6: Cooperation in treatments without observers

	Cooperate		
	All rounds	*First rounds*	*Last game*
Image concern (μ)	-0.36*	-0.73**	-0.78**
	(0.17)	(0.24)	(0.29)
Player female	-0.13	-0.20	-0.63
	(0.35)	(0.38)	(0.39)
Age	0.01	-0.01	-0.02
	(0.01)	(0.02)	(0.02)
Player is student	-0.14	-0.53	-0.41
	(0.54)	(0.57)	(0.63)
Player French	-0.81^{+}	-0.49	-0.68^{+}
	(0.46)	(0.56)	(0.40)
Risk aversion	0.03	0.09	0.08
	(0.09)	(0.08)	(0.09)
Response time	-0.00	0.00	-0.08
	(0.01)	(0.01)	(0.05)
Round fixed effects	Yes	Yes	Yes
Observations	640	120	120
Pseudo R^2	0.07	0.12	0.18

Note: This table reports the estimations of a probit regression of the indicator variable taking the value 1 if the player cooperated, restricting the sample to treatments with no observers. All regressions control for round fixed effects. Column 1 includes all rounds of the prisoner's dilemma game. Column 2 is restricted to the first round of each game. Column 3 restricts the sample to the third game. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Cooperation rates in infinitely repeated games can reflect both an intrinsic level of cooperativeness (due to altruism, for instance), but also beliefs about how likely others are to cooperate.[22] We favor the interpretation based on altruism since the result seems even stronger when we restrict the sample to the last game (column 3) where beliefs should be less heterogeneous since learning will already have occurred during earlier games. This suggests an initial picture where image concerned individuals appear less altruistic: they are more concerned with themselves and the impression they give to others – a result that should be explored in later work.

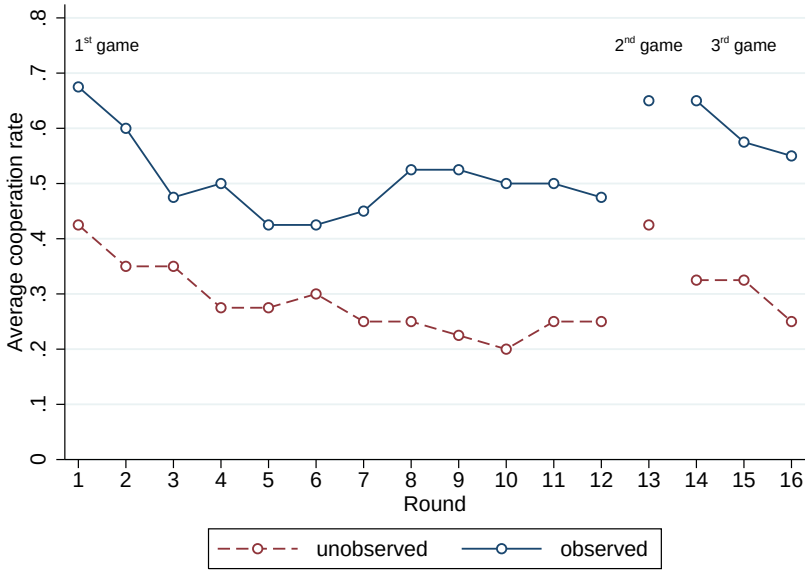
2.4.3 The effect of observers in infinitely repeated prisoner's dilemma

On average, individuals are more likely to follow the social norm in the repeated prisoner's dilemma when they are observed. Comparing the average behavior in treatments with observers to those without, we find that cooperation is more likely in the presence of observers. Figure 2.4 plots cooperation rates for treatments with and without observers over the course of the three

[22]Whether an individual cooperated in the very first round can be seen as a measure for his base level of cooperativeness. In most experiments on infinitely repeated games, this measure significantly increases cooperation in the remainder of the games (see for instance Dal Bó and Fréchette (2011)).

games played by participants. The presence of an observer increases cooperation in all rounds of all games. In treatment without observers, the average cooperation rate is 0.3 while it is 0.53 with observers. The difference is significant ($t = 3.04$, adjusting for participant clusters). There seems to be no difference in the decay of cooperation over time or in the intensity of the restart effect.

Figure 2.4: Cooperation rates in the infinitely repeated prisoner's dilemma



Note: This graph plots average cooperation rates in the infinitely repeated prisoner's dilemma game across the cumulative rounds played in 3 games.

In the presence of observers participants are also more likely to follow the second norm that appears to influence ratings, the consistency in cooperation. When we restrict the analysis to rounds where the player cooperated in the previous round, the rate of cooperation is 0.66 without observers and 0.77 with observers and this difference is (weakly) significant ($t = 1.70$, adjusting for participant clusters). When on the contrary, we focus on the case where the player deviated in the previous round, the rate of cooperation falls to 0.11 with observers and 0.15 without observers, a difference no longer statistically significant ($t = 1.02$, adjusting for participant clusters). Being observed pushes players to follow the social norm more closely, even along subtle dimensions of the norm. This is coherent with the literature that analyzes behavior in public good games when observed (Andreoni and Petrie (2004)) and other experimental games played with observers.

2.4.4 Further validation of our image concern measure

Finally, we use the effect of observers in the repeated prisoner's dilemma to provide further validation of our image concern measure: More image concerned individuals become more prone to behave according to the social norm when observed. As reported in Table 2.7, image concerned participants react more than others to the presence of observers. The interaction

term between image concern and being observed in column 3 positive and significant. However, this is not the case immediately and only becomes apparent in later games when participants have learned the dynamics of the game (column 4).

Table 2.7: Cooperation and image concern

	Cooperate			
	Unobs. sample	Observed sample	Full sample	Full sample
Image concern (μ)	-0.78**	0.11	-0.62*	-0.57*
	(0.29)	(0.16)	(0.27)	(0.23)
Observed			1.06**	0.89**
			(0.27)	(0.26)
Observed $\times$ μ			0.64*	0.31
			(0.29)	(0.27)
Control variables	Yes	Yes	Yes	Yes
Round fixed effects	Yes	Yes	Yes	Yes
Games/Rounds	Last game	Last game	Last game	First rounds
Observations	120	120	240	240
Pseudo R^2	0.18	0.26	0.25	0.17

Note: This table reports the estimations of a probit regression of the indicator variable taking the value 1 if the player cooperated. The image concern measure has been divided by its standard deviation for easier interpretation of the magnitudes. Except for column 4, only observations from the last game are included. Column 1 includes only the observations from sessions where the game was played without observers and column 2 those where the players are observed. We control for socioeconomic characteristics, attitudes towards risk, response time and whether the player knew the observer. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

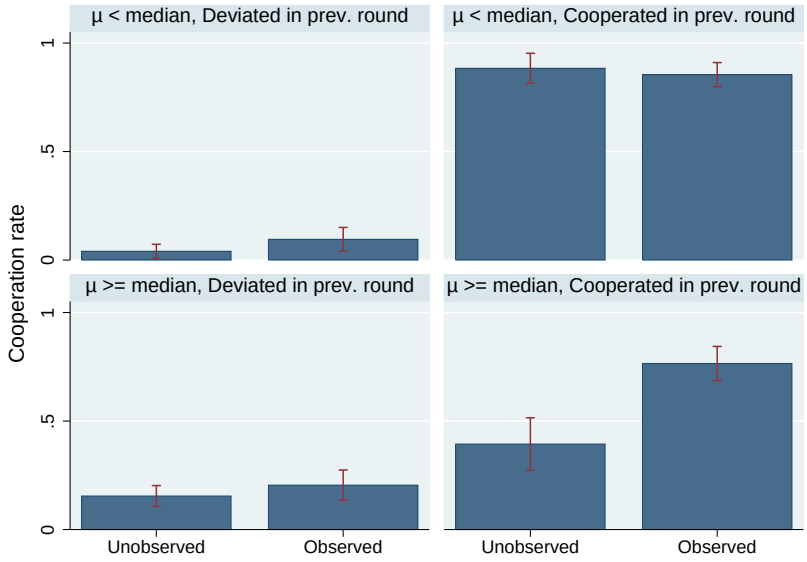
Similarly, under observation, image concerned individuals also react more strongly to the social norm of consistency in cooperation, which was highly rated by observers. This can be seen graphically in Figure 2.5, where the first row is for individuals who are not very image concerned (below median of 10) and the second row is for those who are very concerned (above median): For individuals who cooperated in the previous round, the effect on cooperation of being observed is much stronger when the individual is image concerned than when not.

These results are confirmed in Table 2.8. In column 1 we restrict the sample to non-observed sessions and we see that more image concerned individuals tend to cooperate less after a round where they cooperated than less image concerned individuals. However we see in column 2 that for observed sessions, this effect disappears. Image concerned individuals seem to take into account the fact that this lack of consistency in cooperation is badly perceived by observers. This is confirmed in column 3 where the interaction term between the level of image concern, cooperation in previous round and the fact of being observed is positive and significant.

2.5 Discussion

We have presented in this chapter a novel experimental game to measure image concern, validated the measure and presented initial facts about image concerned individuals. We now

Figure 2.5: Cooperation rates depending on past behavior and presence of observers



Note: This figure contains average cooperation rates in the infinitely repeated prisoner's dilemma game along with 95% confidence intervals. The top and bottom row report averages for participants with image concern levels below and above the median of 10, respectively. The left column contains cases in which the player deviated in the previous round, the right column corresponds to previous cooperation.

discuss some of the possible issues that might be raised about this game.

Experimenter as observer

One might worry that the dictator is under the impression that he is being observed, not only by the observer in the game, but also by the experimenter. If the dictator believes that the experimenter can see how much he pays to be anonymous and if he thinks the experimenter will adversely interpret payments for anonymity, it could imply that image concerned individuals could be less inclined to give than if the experimenter was not present.

We took several precautions to limit this potential problem. First, we clearly told the participants that they would remain anonymous from the point of view of the experimenter. They were told that the photos would be deleted at the end of the session, and that we would of course preserve their anonymity while we conduct the analysis. Furthermore, while the picture of the observer was always visible on the screen of the dictators, the experimenter was not visible during the experiment.

However, even if despite these precautions subjects were still influenced by the experimenter, this would have little impact on the results presented here. It would only decrease the variance in the answers but not change the ranking of individuals in terms of μ.[23] The fact that we find a high degree of heterogeneity in the population, suggests that even if this effect were present,

[23]Unless of course there are two dimensions of image concern that can both vary across the population: being concerned about the experimenter's perception of the level of generosity and being concerned about the experimenter's perception about trying to hide one's true type.

Table 2.8: Cooperation as a function of player's previous round action

	Cooperate		
	Unobs. sample	Observed sample	Full sample
Image concern (μ)	0.19	-0.89^{+}	-0.11
	(0.37)	(0.49)	(0.20)
Player cooperated in previous round	1.00^{+}	2.94**	1.61**
	(0.53)	(0.76)	(0.31)
Observed			1.00**
			(0.32)
Image con. × cooperated in previous round	-2.78**	2.01**	-1.48**
	(0.95)	(0.75)	(0.45)
Image con. × coop. in prev. round × observed			2.45**
			(0.61)
Controls	Yes	Yes	Yes
Game × round FE	Yes	Yes	Yes
Observations	80	80	160
Pseudo R^2	0.52	0.50	0.48

Note: This table reports the estimations of a probit regression of the indicator variable taking the value 1 if the player cooperated. The image concern measure has been divided by its standard deviation for easier interpretation of the magnitudes. Only observations from the last game are included. Column 1 includes only the observations from sessions where the game was played without observers and column 2 those where the players are observed. We control for socioeconomic characteristics, attitudes towards risk, response time and whether the player knew the observer. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

it would not be large.

Shy individuals

Another potential worry could be that our experiment just measures how shy an individual is: Some people might be ready to pay for their picture not to be shown, even if no actions are revealed. We believe that the three different validation methods of our measurement that we presented in section 2.3.2 and 2.4.4 reject this alternative story. There is no reason for shy people to systematically answer our survey question differently or to follow the social norm when they are observed in the infinitely repeated game. It might nevertheless be interesting to run an alternative experiment where the decision would be to pay not to have a picture revealed without any actions being attached to it.

Different sensitivities depending on the characteristic

In our experiment, the inferences that the observer can make about the characteristics of an individual is restricted to the level of altruism. It is conceivable that individuals might care differently about their image depending on the characteristic at stake. Using the terminology of the model, for different characteristics k, the γ_k^i might be different (see equation (2.2)). We cannot test this directly, but we found that those individuals identified as being image

concerned in our game also react to observers in situations where their actions reveal different things than altruism: For instance, they respect the second norm of behavior in the repeated prisoner's dilemma more, a norm that has less to do with altruism. This question nevertheless remains one that should be examined in future research.[24]

Observer characteristics

The dictator's willingness to pay to remain anonymous in the image concern games seems not to be influenced much by the observer's characteristics – at least in our small laboratory sample. We investigate the four features that participants could easily infer from the pictures of observers: gender, approximate age, nationality (French vs. non-French) and facial expression. Table 2.9 presents the results. Column 1 shows that none of these features have any influence on average. The fact that the observer smiles has a slight positive impact that tends not to be significant.

Columns 2 to 4 introduce interaction terms to understand the role of observers in more detail. While we find no significant effect for age and gender, Non-French dictators appear to be less concerned about their image when they are observed by another non-French participant and slightly but not significantly more when observed by a French.[25]

The fact that bids for anonymity seem not to be driven by the observer's characteristics suggests that the image concern game can be used in a wide range of contexts inside and outside the laboratory. However, experimenters should account for the differential response participants may have when observed by a minority member.

Multiple hypothesis testing

A growing body of literature highlights the risk of false discoveries when testing multiple hypotheses within the same experiment (see for instance List et al. (2016)). Indeed, when analyzing several subgroups and correlating multiple survey questions and experimental outcomes, one may worry that rejecting null hypotheses at traditional significance levels leads to an elevated risk of type I error. We thus conduct additional robustness checks to minimize the risk that any of the results presented above are false discoveries. Following an approach proposed by Benjamini and Hochberg (1995), we collect all the hypotheses tested in the regressions and apply their method to control the False Discovery Rate (FDR), defined as the share of false rejections among all rejections. The resulting corrected p-values (q-values) associated with all the hypotheses investigated in this chapter are reported in Table 2.12 in the appendix. They should be interpreted as the minimum FDR for rejection of the associated null hypothesis of no effect.

[24] Bracha and Vesterlund (2013), for instance, describe an experiment separating generosity signaling from income signaling.

[25] We use the variable coding nationality as French or not, as a proxy for the race of individuals, which we cannot ask directly. Most of the non-French are nationals from North African countries.

Despite the fact that this method is much more conservative than regular hypothesis testing, our main results still hold – even if some of them require lower thresholds to rejection. In particular, all three validations of our image concern measure presented above are confirmed.

2.6 Conclusion

This chapter proposes an experimental procedure to measure individual sensitivity to image concern, validates the measure and starts exploring determinants and consequences of this underexplored dimension of preferences. It opens the way for future research on the topic.

As a first step in this direction, we are able to document two patterns: First, the extent to which individuals are concerned with their social image is highly heterogeneous and does not depend on the observer characteristics. As a notable exception, we provide evidence that minority members might be more concerned with the image that they project to members of the majority group. Second, we show that image concerned individuals are less cooperative in classic infinitely repeated prisoner's dilemma games, but that cooperation rates increase up to the average when the same game is played under the scrutiny of a third party observer. This suggests that for these individuals, norm compliance is essentially linked to the salience of their actions. Exploring the heterogeneity of image concerns further and linking it to established measures of personality traits such as the "Big Five" remains a promising direction for future research.

Due to the nature of the concept to be measured, the game is of course less portable than other games aimed at measuring social preferences, such as the trust game or the dictator game and might be more difficult to run in a remote field environment. Variations in the setup can, however, be used. For instance, the process of taking pictures could be eliminated by asking the dictator to stand up if he loses anonymity. There are also ways to implement the BDM without a computer terminal. Finally, in settings where retaliation outside the laboratory is unlikely, experimenters can consider to use the recipient as an observer, rather than adding a third party observer. Overall, we attempted to ensure simplicity of the setup while preserving robustness of the measure and avoiding confounding factors such as other social preferences.

Table 2.9: Role of observers

	Image concern			
Observer female	0.16		-0.16	-0.14
	(2.01)		(1.99)	(2.01)
Observer age	-0.04	-0.03		-0.04
	(0.07)	(0.07)		(0.07)
Observer French	0.98	1.31	0.64	
	(2.98)	(2.89)	(3.13)	
Observer smiles	5.30	5.08	5.01	4.46
	(7.62)	(7.47)	(7.69)	(7.63)
Observer frowns	6.04	6.37	6.26	6.25
	(7.71)	(7.57)	(7.93)	(7.28)
Picture blurry	-4.35	-4.07	-4.06	-3.88
	(6.62)	(6.43)	(6.85)	(6.58)
Male × observer female		3.19		
		(3.38)		
Female × observer male		5.47		
		(3.44)		
Female × observer female		2.96		
		(3.02)		
>24y old × observer ≤24y old			-2.10	
			(2.60)	
≤24y old × observer >24y old			-5.50	
			(3.41)	
≤24y old × observer ≤24y old			-3.34	
			(3.44)	
Non-French × observer French				13.42**
				(4.33)
French × observer non-French				9.32*
				(3.71)
French × observer French				8.17**
				(2.88)
Observations	367	367	367	367
R^2	0.160	0.168	0.166	0.172

Note: This table presents regressions of the bid for anonymity on dictators' and observers' characteristics. All regressions control for round fixed effects and for outcomes of the previous round, though the results are qualitatively unchanged when these controls are omitted. Dictators that knew their observer are omitted from the regression but results remain similar when this is not done. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

References

Algan, Yann, Yochai Benkler, Emeric Henry, and Jérôme Hergueux. 2014. "Social Motives and the Organization of Production : Experimental Evidence from Open Source Software." *Working paper.*

Algan, Yann, Yochai Benkler, Mayo Fuster Morell, and Jérôme Hergueux. 2013. "Cooperation in a peer production economy - experimental evidence from Wikipedia." *Working Paper.*

Ali, Nageeb S., and Roland Bénabou. 2016. "Image Versus Information: Changing Societal Norms and Optimal Privacy." *NBER Working Paper.*

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77 (5): 1607–1636.

Andreoni, James, and Ragan Petrie. 2004. "Public goods experiments without confidentiality: A glimpse into fund-raising." *Journal of Public Economics* 88 (7-8): 1605–1623.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99 (1): 544–555.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring utility by a Single Response Sequential Method." *Systems Research and Behavioural Science* 9 (3): 226–232.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652–1678.

———. 2012. "Laws and Norms." *IZA Discussion Paper* 6290:1–44.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society B* 57 (1): 289–300.

Bracha, Anat, and Lise Vesterlund. 2013. "How Low Can You Go ? Charity Reporting When Donations Signal Income and Generosity." *Working paper.*

Bursztyn, Leonardo, and Robert Jensen. 2015. "How does peer pressure afect educational investments?" *Quarterly Journal of Economics* 130 (3): 1329–1367.

Carpenter, Jeffrey, and Caitlin Knowles Myers. 2010. "Why volunteer? Evidence on the role of altruism, image, and incentives." *Journal of Public Economics* 94 (11-12): 911–920.

Charness, Gary, Luca Rigotti, and Aldo Rustichini. 2007. "Individual behavior and group membership." *American Economic Review* 97 (4): 1340–1352.

Costa, Paul, and Robert McCrae. 2008. "The revised NEO personality inventory (NEO-PI-R)." In *The SAGE Handbook of Personality Theory and Assessment,* 179–198.

Dal Bó, Pedro. 2005. "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games." *American Economic Review* 95 (5): 1591–1604.

Dal Bó, Pedro, and Guillaume R Fréchette. 2011. "The Evolution of Cooperation in Infinitely Repeated Games." *American Economic Review* 101 (1): 411–429.

Dana, Jason, Daylian M. Cain, and Robyn M Dawes. 2006. "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." *Organizational Behavior and Human Decision Processes* 100 (2): 193–201.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness." In *Economic Theory,* 33:67–80. 1.

Dellavigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." *Quarterly Journal of Economics* 127 (1): 1–56.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association* 9 (3): 522–550.

Ekström, Mathias. 2012. "Do watching eyes affect charitable giving? Evidence from a field experiment." *Experimental Economics* 15 (3): 530–546.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of lncentive Theory." *American Economic Review* 98 (3): 990–1008.

———. 2011. "Conspicuous generosity." *Journal of Public Economics* 95, nos. 9-10 (October): 1131–1143.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and Human Behavior* 25 (2): 63–87.

Fehr, Ernst, Bernhard Von Rosenbladt, Gert G. Wagner, Urs Fischbacher, and Jürgen Schupp. 2003. "A Nation-Wide Laboratory : Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys." *IZA Discussion Paper,* no. 715.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a segmented society: An experimental approach." *Quarterly Journal of Economics* 116 (1): 351–377.

Fudenberg, Drew, David G. Rand, and Anna Dreber. 2012. "Slow to anger and fast to forgive: Cooperation in an uncertain world." *American Economic Review* 102 (2): 720–749.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *Quarterly Journal of Economics* 115 (3): 811–846.

Lacetera, Nicola, and Mario Macis. 2010. "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme." *Journal of Economic Behavior and Organization* 76 (2): 225–237.

List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." *NBER Working Paper.*

Peysakhovich, Alexander, and David G. Rand. 2015. "Habits of virtue: creating norms of cooperation and defection in the laboratory." *Management Science* 62 (3): 631–647.

Rege, Mari, and Kjetil Telle. 2004. "The impact of social approval and framing on cooperation in public good situations." *Journal of Public Economics* 88 (7-8): 1625–1644.

Riyanto, Yohanes E, and Jianlin Zhang. 2015. "Putting a price tag on others' perceptions of us." *Experimental Economics* 19 (2): 480–499.

Samek, Anya Savikhin, and Roman M. Sheremeta. 2014. "Recognizing contributors: an experiment on public goods." *Experimental Economics* 17 (4): 673–690.

Sapienza, Paola, Anna Toldra-Simats, and Luigi Zingales. 2013. "Understanding trust." *Economic Journal* 123 (573): 1313–1332.

Sutter, Matthias, Peter Lindner, and Daniela Platsch. 2009. "Social norms, third-party observation and third-party reward." *Working paper.*

Appendices

2.A Proofs

2.A.1 Proof of Proposition 1

Equilibrium bid b^* increasing in image concern

We first show that the equilibrium bid b^* is increasing in image concern μ. Under the assumption that the cost functions are separable, individual i chooses the amount to bid in the dictator lottery X and the amount to bid for anonymity b to maximize

$$X\left[v^i(1)+\mu R_1^*\right]+(1-X)\left[v^i(0)+(1-b)\mu R_0^*-bc_2(b)\right]-c_1(X)$$

R_1^* is the image, in equilibrium, when the lottery is a success while R_0^* is the image in case of a failure.

In an interior equilibrium, using the notation $v=v^i(1)-v^i(0)$, the first order conditions relative to X and b yield:

$$c_1^{'}(X^*)=v+\mu R_1^*-[(1-b^*)\mu R_0^*-bc_2(b^*)] \tag{2.3}$$

$$b^*c_2^{'}(b^*)+c_2(b^*)=-\mu R_0^* \tag{2.4}$$

Denoting L the outcome of the lottery we have by the law of iterated expectations

$$P[L=1]E[v|L=1]+P[L=0]E[v|L=0]=v_0$$

i.e

$$P[L=1]\left(E[v|L=1]-v_0\right)+P[L=0]\left(E[v|L=0]-v_0\right)=0$$
$$\Leftrightarrow P[L=1]R_1^*+P[L=0]R_0^*=0$$

Thus, since $E[v|L=1]>E[v|L=0]$, $R_1^*>0$ and $R_0^*<0$, i.e. the fact that the lottery is a success makes the observer update positively her beliefs on the amount transferred by the individual. Since $R_0^*<0$ and the cost function is increasing and convex, condition (2.4) implies that b^* is increasing in μ.

Results 1 and 2

We now establish results 1 and 2. These results are derived under the assumption of separability. The first order condition (2.4) directly implies that b^* is independent of v.

Furthermore, taking the total derivative of condition (2.3), we have

$$\frac{\partial X^*}{\partial \mu} = \frac{-(1-b^*)R_0^* + R_1^* + \frac{\partial b^*}{\partial \mu}\left(\mu R_0^* + c_2(b^*) + b^* c_2'(b^*)\right)}{c_1''(X^*)}.$$

Using condition (2.4), this implies

$$\frac{\partial X^*}{\partial \mu} = \frac{-(1-b^*)R_0^* + R_1^*}{c_1''(X^*)}.$$

So this establishes result 2, since $R_0^* < 0$, $R_1^* > 0$ and $c(\cdot)$ is convex.

2.A.2 Non-separable cost functions

Proposition 2: If the cost function is not separable, a sufficient condition for b^* to be increasing in μ is that

$$B_2 \equiv c^{''}(X^*) + (1 - X^*)b^* \left(c^{''}(X^* + b^*) - c^{''}(X^*)\right) + 2b^* \left(c^{'}(X^*) - c^{'}(X^* + b^*)\right) < 0.$$

Furthermore, if $c(x) = e^{\alpha x}$, with $\alpha < 1$, there exists $\bar{b}$ such that, if $b > \bar{b}$, the condition $B_2 \leq 0$ is satisfied.

Proof

In the case where the cost functions are not separable, we denote c the strictly increasing and convex cost function. In this case, Individual i thus chooses X and b to maximize

$$X\left[v^i(1) + \mu R_1^* - c(X)\right] + (1 - X)\left[v^i(0) + (1 - b)\mu R_0^* - bc(b + X) - (1 - b)c(X)\right] \tag{2.5}$$

The first order condition with respect to X can be expressed as

$$\begin{aligned} & c^{'}(X^*)\left(X^* + (1 - X^*)(1 - b^*)\right) + (1 - X^*)b^*c^{'}(X^* + b^*) && (2.6)\\ = \; & v + \mu R_1^* - \left[(1 - b^*)\mu R_0^* - b^*c(X^* + b^*) + b^*c(X^*)\right] && (2.7)\end{aligned}$$

The first order condition with respect to b can be expressed as:

$$b^*c^{'}(X^* + b^*) + c(X^* + b^*) - c(X^*) = -\mu R_0^* \tag{2.8}$$

Taking the total derivative of condition (2.8) with respect to μ gives

$$\frac{\partial b^*}{\partial \mu} A_1 + \frac{\partial X^*}{\partial \mu} B_1 = -R_0^* \tag{2.9}$$

where

$$\begin{aligned} A_1 &= c^{'}(X^* + b^*) + b^*c^{''}(X^* + b^*) + c^{'}(X^* + b^*) \\ &= 2c^{'}(X^* + b^*) + b^*c^{''}(X^* + b^*) > 0 \end{aligned}$$

and

$$B_1 = b^*c^{''}(X^* + b^*) + c^{'}(X^* + b^*) - c^{'}(X^*) \geq 0$$

Taking the total derivative of condition (2.7) with respect to μ can be expressed as:

$$\frac{\partial b^*}{\partial \mu} A_2 + \frac{\partial X^*}{\partial \mu} B_2 = R_1^* - (1 - b^*) R_0^* \tag{2.10}$$

where

$$\begin{aligned} A_2 &= -c^{'}(X^*)(1 - X^*) + (1 - X^*)c^{'}(X^* + b^*) + (1 - X^*)b^* c^{''}(X^* + b^*) \\ &\quad -\mu R_0^* - c(X^* + b^*) - b^* c^{'}(X^* + b^*) + c(X^*) \end{aligned}$$

Using the condition (2.8) we have $-\mu R_0^* - c(X^* + b^*) + c(X^*) = b^* c^{'}(X^* + b^*)$, so that

$$A_2 = (1 - X^*)\left(c^{'}(X^* + b^*) - c^{'}(X^*)\right) + (1 - X^*)b^* c^{''}(X^* + b^*) > 0$$

and

$$\begin{aligned} B_2 &= c^{''}(X^*)\left(X^* + (1 - X^*)(1 - b^*)\right) + c^{'}(X^*)\left(1 - (1 - b^*)\right) - b^* c^{'}(X^* + b^*) \\ &\quad +(1 - X^*)b^* c^{''}(X^* + b^*) - b^* c^{'}(X^* + b^*) + b^* c^{'}(X^*) \end{aligned}$$

i.e.

$$B_2 = c^{''}(X^*) + (1 - X^*)b^*\left(c^{''}(X^* + b^*) - c^{''}(X^*)\right) + 2b^*\left(c^{'}(X^*) - c^{'}(X^* + b^*)\right) \tag{2.11}$$

Combining conditions (2.10) and (2.9), we have

$$\frac{\partial b^*}{\partial \mu}\left[A_1 - \frac{B_1}{B_2} A_2\right] = -R_0^* - \frac{B_1}{B_2}\left(R_1^* - (1 - b^*)R_0^*\right)$$

i.e

$$\frac{\partial b^*}{\partial \mu} = -\frac{B_2}{A_1 B_2 - A_2 B_1} R_0^* - \frac{B_1}{A_1 B_2 - A_2 B_1}\left(R_1^* - (1 - b^*)R_0^*\right) \tag{2.12}$$

Given that $R_0 < 0$, $R_1 > 0$, $A_1 > 1$, $A_2 > 0$ and $B_1 \geq 0$, a sufficient condition for $\frac{\partial b^*}{\partial \mu} > 0$ is $B_2 < 0$.

We know turn to the second part of the Proposition. When $c(x) = e^{\alpha x}$, we have:

$$\begin{aligned} B_2 &= \alpha^2 e^{\alpha X} + (1 - X)b\alpha^2\left(e^{\alpha(X+b)} - e^{\alpha X}\right) + 2b\alpha\left(e^{\alpha X} - e^{\alpha(X+b)}\right) \\ &= \alpha e^{\alpha X}\left[\alpha + b\left(e^{\alpha b} - 1\right)\left((1 - X)\alpha - 2\right)\right] \end{aligned}$$

Given that $\alpha \leq 1$, we have $(1 - X)\alpha - 2 < 0$. Furthermore, B_2 is decreasing in b and at the limit, for $b = 1$, we have that B_2 is proportional to $\alpha + (e^{\alpha} - 1)\left((1 - X)\alpha - 2\right)$ which is negative for $\alpha < 1$. Thus, by continuity, there exists $\bar{b}$ such that, if $b > \bar{b}$, $B_2 < 0$ and $\frac{\partial b^*}{\partial \mu} > 0$.

2.B Experimental instructions

We include the instructions for the experiment (translated from French into English). The experiment contained three parts that were handed out separately:

1. the image concern game
2. the repeated prisoner's dilemma game, with or without observers [26]
3. a survey questionnaire

At the end of each part, a short questionnaire was included so that participants could check their own understanding of the instructions and ask questions if necessary. The correct answers were then read out loud to ensure the participants understood everything correctly.

[26] We include only the instructions for the observed version here, the unobserved version used essentially the same instructions but without any mention of observers.

WELCOME!

Thank you for taking part in this experiment. Please turn off your cell phones now before the experiment begins.

You can earn money over the course of this experiment. The amount you will earn depends on the decisions that you and the other participants take during the experiment. All the information that you will share during the experiment as well as the amount you made will remain strictly confidential and anonymous. Your photo will be deleted at the end of the experimental procedure.

For obvious scientific reasons, you are not allowed to talk during the experiment. We have no choice but to ask participants who break this rule to leave the laboratory without receiving the gains they may have made during the experiment.

This experiment consists of 3 completely independent parts. You will receive new instructions prior to each part.

INSTRUCTIONS FOR PART 1

In the beginning of this experiment, a random draw determines your role: You will be either player A, B or C. Your role remains the same for the entire part 1 of the experiment. A message on the screen will inform you of the role you are assigned.

Part 1 of the experiment consists of 4 rounds, but your role will remain the same throughout part 1. At the end of the experiment, one of the 4 rounds will be selected at random and determine your gains for part 1 of the experiment. Each round thus has a chance of one in 4 to determine the amount you will make in part 1 of the experiment.

Procedure for any given round

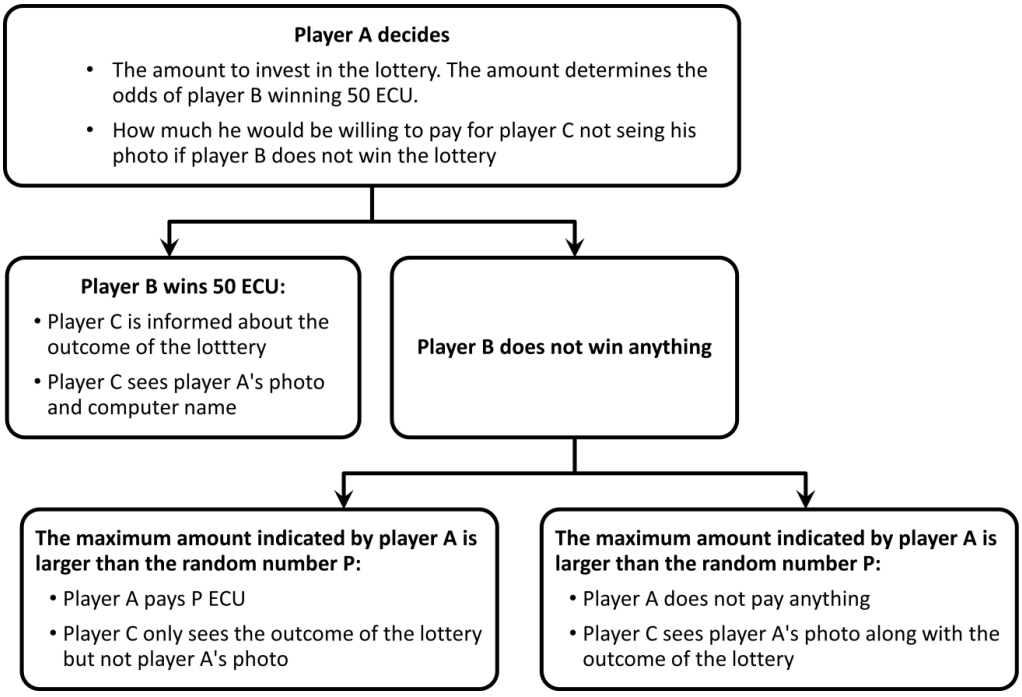
At the beginning of each round, groups of three players are formed at random. Each of the groups comprises a player A, a player B and a player C. Players A and B are part of only one group, whereas players of type C are assigned to two groups at a time.

You will change groups after each round and two participants will never be in the same group twice throughout the experiment, including both part 1 and part 2 of the experiment. The experimental procedure followed in each round is the following:

1. Player A receives 100 ECU. Player C receives 40 ECU. Player B does not receive anything.
2. Player A can decide to invest money in a lottery. The gains from this lottery will benefit player B only. For each ECU invested in the lottery, the chances of player B winning a fixed sum of 50 ECU increase by 1%. Thus, if player A invests X ECU, player B has a chance of X in 100 to get 50 ECU. Whatever the outcome of the lottery, player A will not recover the amount invested in the lottery.
3. Before the lottery draw is carried out, player A has to indicate the **maximum amount he would be willing to pay for player C not to see his photo in case the result of the lottery is negative** (i.e player B receives 0 ECU).
4. The lottery draw is carried out at the odds determined by the choices of player A in step 2. Depending on the outcome of the lottery, player B receives 50 ECU or nothing at all.

5. Player C is informed of the outcome of the lottery but not of the amount invested by player A. He only knows whether player B received 0 or 50 ECU.
   - **Moreover, if player B wins the lottery, player C sees the photo and computer name of player A**
   - **If player B does not win the lottery (does not receive 50 ECU), the maximum amount indicated by player A in step 3 is used.** What follows now is making sure that player A has interest to honestly report the maximum amount in step 3: a random number P between 0 and 100 is drawn
     - If P is larger than the maximum amount indicated by player A, player A's photo is displayed on player C's screen but player A does not pay anything.
     - If P is smaller than the maximum amount indicated by player A, the photo is not displayed and player A pays P.

The round ends with the completion of step 5. The procedure followed in any given round is summed up in the following graph:



After each round, you will be informed about the gains you made during that round. Then, new groups of three are formed at random and a new round begins. There are 4 rounds in total. The groups change after each round so that you never encounter another participant more than one time throughout the experiment.

Computing your gains

At the end of the experiment, one of the 4 rounds in this part of the experiment will be selected at random. The outcome of the selected round will determine your gains for this part of the experiment. The conversion rate between ECU and Euro is: 10 ECU = 1 Euro.

In addition, you will receive a flat amount of 4 Euros and the gains you may make during the second part of the experiment.

THANKS FOR COMPLETING THIS QUESTIONNAIRE

1. Your role remains the same throughout this part of the experiment.
   ☐ True
   ☐ False

2. The groups remain the same throughout this part of the experiment.
   ☐ True
   ☐ False

3. The odds of player B winning the lottery are determined by player A.
   ☐ True
   ☐ False

4. Player C always sees player A's photo and computer name.
   ☐ True
   ☐ False

5. Player B can see player A's photo and computer name.
   ☐ True
   ☐ False

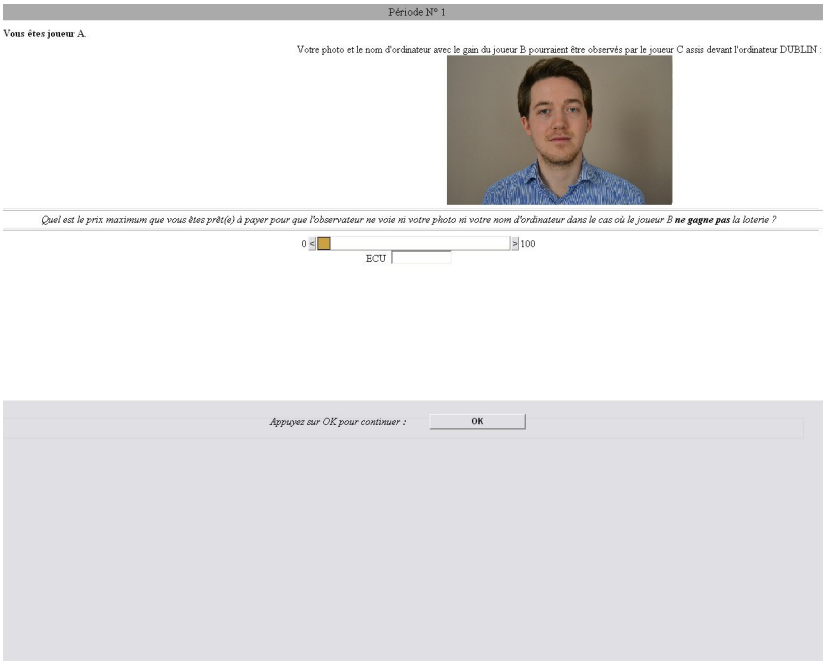
6. Player C' gains depend on the other players' decisions.
   ☐ True
   ☐ False

7. Player C knows the exact amount player A invests in the lottery.
   ☐ True
   ☐ False

8. Your gains from this part of the experiment are determined by only one of the four rounds.
   ☐ True
   ☐ False

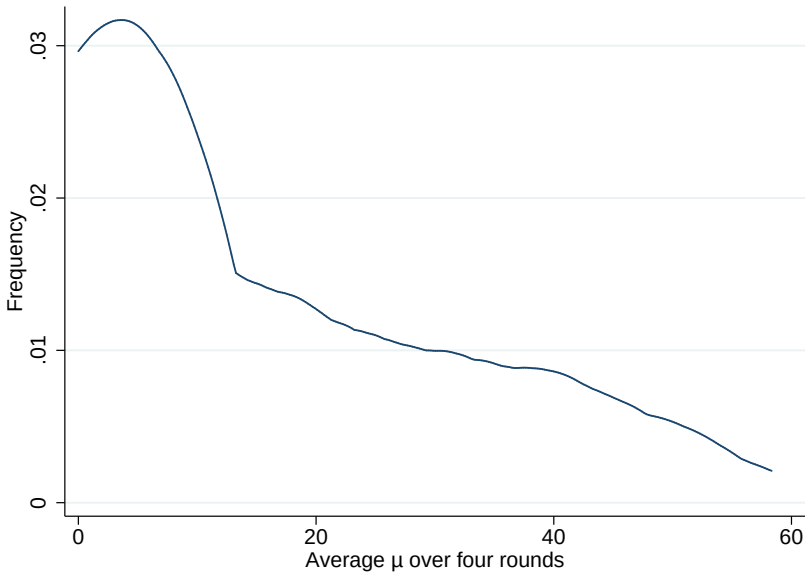
2.C Supplementary tables and figures

Figure 2.6: Screenshot of the user interface for participants in the experiment



Note: Screenshot of the software used in the experiment. This screen reads "You are player A. Your photo and computer name could become visible to the player C sitting at computer 'Dublin' [photo of player C]. What is the maximum price you are willing to pay so that this observer does not see your picture and computer name in case player B does not win the lottery? Click 'OK to continue."

Figure 2.7: Distribution of average bids for anonymity across all rounds



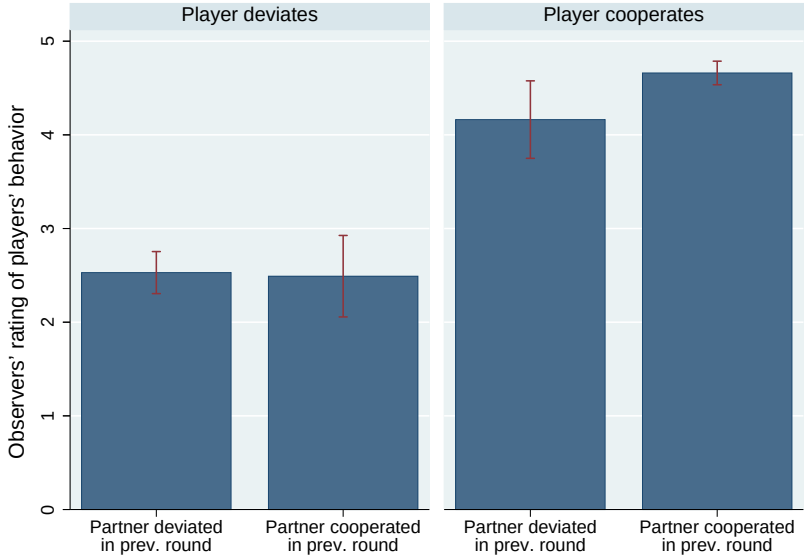
Note: This figure plots a kernel density estimate of participants' image concern as measure by their average bid for anonymity across all four rounds (Epanechnikov, bandwitdh = 5.9).

Table 2.10: Observers' ratings in the infinitely repeated prisoner's dilemma game

	Top rating		
Cooperate	1.63**	1.32**	1.08**
	(0.26)	(0.35)	(0.35)
Player French	3.34**	3.45**	3.35**
	(0.62)	(0.62)	(0.60)
Observer French	4.18**	4.10**	4.16**
	(0.34)	(0.37)	(0.34)
Observer French × player French	-3.47**	-3.54**	-3.44**
	(0.60)	(0.60)	(0.57)
Player is student	0.70+	0.73+	0.69+
	(0.38)	(0.38)	(0.37)
Player smiles	0.81*	0.68+	0.64+
	(0.41)	(0.40)	(0.38)
Deviate after partner cooperated		0.18	
		(0.29)	
Cooperate after partner cooperated		0.48+	
		(0.26)	
Deviate after player cooperated			-0.09
			(0.24)
Cooperate after player cooperated			0.72*
			(0.34)
Observations	520	520	520
Pseudo R^2	0.344	0.351	0.356

Note: This table reports the estimations of a probit regression of the indicator variable taking the value 1 if the observer gave a rating of 5 on the player's characteristics and behaviors. We control for players' and observers' age and gender as well as attitudes towards risk. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Figure 2.8: Rating of behavior by observer depending on past choices of partner



Note: This figure reports average ratings of players' decisions in the infinitely repeated prisoner's dilemma game by outside observers. 95% confidence intervals in red.

Table 2.11: Cooperation and image concern using survey question

	Cooperate			
	Unobs. sample	Obs. sample	Full sample	Full sample
Do not want to appear selfish	-0.26	-0.47	-0.17	-0.11
	(0.33)	(0.40)	(0.35)	(0.40)
Observed			1.06*	0.59
			(0.43)	(0.42)
Observed × not appear selfish			-0.26	0.41
			(0.50)	(0.52)
Control variables	Yes	Yes	Yes	Yes
Round fixed effects	Yes	Yes	Yes	Yes
Games/Rounds	Last game	Last game	Last game	First rounds
Observations	120	120	240	240
Pseudo R^2	0.10	0.27	0.23	0.13

Note: This table reports the estimations of a probit regression of the indicator variable taking the value 1 if the player cooperated. "Do not want to appear selfish" reports a 1-5 answer to the question "It is important to me not to be perceived as selfish". Except for column 4, only observations from the last game are included. Column 1 includes only the observations from sessions where the game was played without observers and column 2 those where the players are observed. We control for socioeconomic characteristics, attitudes towards risk, response time and whether the player knew the observer. Standard errors in parentheses are clustered at the individual level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 2.12: Result of Benjamini-Hochberg (1995) multiple testing correction

Table	Coefficient	Point estimate	p-value	q-value
3 (1)	Female	-0.45	0.91	0.95
	Age	0.12	0.56	0.70
	In a relationship	2.22	0.57	0.71
	Student	-8.01	0.20	0.41
	French	-3.59	0.55	0.70
	Knows the observer	5.39	0.43	0.65
4 (1)	Image concern (μ)	0.50	0.01*	0.03*
4 (2)	Image concern (μ)	7.98	<0.01*	0.03*
5 (1)	Cooperate	2.94	<0.01*	<0.01*
5 (3)	Cooperate	1.78	0.01*	0.03*
	Cooperate after player cooperated	1.60	0.02*	0.07*
6 (3)	Image concern (μ)	-0.77	0.01*	0.03*
7 (3)	Observed	1.06	<0.01*	<0.01*
	Image concern (μ)	-0.62	0.02*	0.07*
	Image concern× observed	0.64	0.03*	0.09*
8 (3)	Image concern (μ)	-0.11	0.59	0.72
	Cooperated in previous period	1.61	<0.01*	<0.01*
	Image × coop. previous round	-1.48	<0.01*	0.01*
	Image × coop. prev. × observed	2.45	<0.01*	<0.01*
	Observed	1.00	<0.01*	0.02*
9 (1)	Observer female	0.16	0.94	0.96
	Observer age	-0.04	0.55	0.70
	Observer french	0.98	0.74	0.83
9 (2)	Male × observer female	3.19	0.35	0.61
	Female × observer male	5.47	0.11	0.27
	Female × observer female	2.96	0.33	0.61
9 (3)	>24y old × observer <=24y old	-2.10	0.42	0.65
	<=24y old × observer >24y old	-5.50	0.11	0.26
	<=24y old × observer <=24y old	-3.34	0.33	0.61
9 (4)	Non-French × observer French	13.42	<0.01*	0.02*
	French × observer non-French	9.32	0.01*	0.05*
	French × observer French	8.17	0.01*	0.03*

Note: This table reports the coefficients of interest from previous regressions along with their associated p-values and the q-values obtained when adjusting for multiple testing using the Benjamini-Hochberg (1995) procedure. The first column indicates the number and column of the table from which the coefficient is taken. Control variables not reported here but included in the correction procedure. p/q-values ≤ 0.1 are starred.

Chapter 3

Vertical Integration and Foreclosure: Evidence from Production Network Data

This chapter is joint work with Johannes Boehm.

3.1 Introduction

Vertical integration of two firms has the potential to increase their economic efficiency by exploiting synergies in the design, production, and distribution of their goods and services. At the same time, firms may pursue integration as a strategy not only to create competitive advantage, but also to engage in anti-competitive behavior. One such case arises when one of the integrating firms controls access to a bottleneck input, such as access to vital infrastructure or technology. The integrated firm might use its access to the bottleneck to extend or preserve its market power in the upstream markets by refusing to provide rival firms in downstream markets with access to the bottleneck. These firms are said to be foreclosed. While a large theoretical literature investigates the motives for vertical foreclosure[1], empirical evidence of firms using foreclosure as a business strategy is restricted to a few very particular cases[2], not least because vertical relationships are rarely observed. This not only restricts our ability to test the theory, but also limits our understanding of the prevalence of foreclosure in reality. Even less is known about potential strategies to mitigate the effects of being foreclosed.

The empirical prevalence of vertical foreclose is, of course, at least partly determined by competition law and its enforcement. Most of its forms are regarded as violating competition laws in a large range of jurisdictions. In the United States the Sherman and Clayton Antitrust Acts set out limitations to merger activity, and starting with *Terminal Railroad Association v. U.S.* (1912) U.S. courts have established a doctrine on foreclosure. Competition authorities typically issue guidelines on their assessment of vertical mergers to avoid unforeseen restrictions on mergers. At the same time – or perhaps as a consequence – enforcement of these vertical merger laws is relatively rare.[3] With recent work arguing that concentration and market power among US firms increased over the course of the last decades[4], and the finger being pointed at regulatory authorities[5], one is led to ask: is enforcement lax, or is actual foreclosure just very rare? What are the factors determining the prevalence of vertical foreclosure, and how severe are the consequences? How can firms threatened by foreclosure mitigate its impact?

This paper examines the occurrence of vertical foreclosure across a range of industries and countries. We exploit a novel panel dataset on vertical relationships — the network structure of production — between large firms, both in the U.S. and abroad. These data allow us to study whether buyer-seller relationships break following vertical mergers and acquisitions. We show that the breaking of a buyer-seller relationship in response to the supplier vertically integrating downstream is more likely when the downstream merging firm is a competitor of the buyer —

[1] See Rey and Tirole (2007) for an overview. The classic references are Hart and Tirole (1990) and Ordover et al. (1990).

[2] Recent examples include Asker (2016) for the Chicago beer market and Crawford et al. (2018) for the US cable TV industry. Lafontaine and Slade (2007) and Slade (2019) survey this literature.

[3] Salop and Culley (2015) find only 46 vertical enforcement actions in the US over the period 1994–2013.

[4] De Loecker et al. (2018) estimate a rise in average US markups using Compustat and US Census data; Gutiérrez and Philippon (2017) document rising Herfindahl concentration indices in US industries, and Barkai (2016) documents a rise in the profit share of US non-financial corporations.

[5] See Gutierrez and Philippon (2018) and The Economist (2018)

but not when the downstream merging firm is not a competitor of the buyer. Consistent with theories of vertical foreclosure, the former break probability is even higher when there is little competition in the upstream industry. The increased probability of links breaking cannot be explained by common industry-level (or industry-pair-level) shocks to merger activity or the break probability. We find this increased break probability in response to both domestic and cross-border mergers. Similarly, domestic and cross-border relationships are equally likely to break in response to such vertical mergers.

The correlation we find does not immediately imply that vertical market foreclosure is taking place in the population of firms and relationships that we study. Causality could run in the opposite direction: vertical integration could be the *response* to relationships breaking, or to the threat thereof. Alternatively, both integration and links breaking could be caused by unobserved shocks. Finally, the links breaking might not be the the consequence of foreclosure, but might be the consequence of the integrating parties being able to produce the final good at such a low cost that the buyer decides to exit the market (and hence stops purchasing the input).

A series of additional regressions indicates that these explanations are unlikely to account for the findings. To see whether our results stem from reverse causality, we follow Edmans et al. (2012) to construct an instrumental variable for vertical mergers and acquisitions. The variable captures events where investor capital outflows of mutual funds put large downward pressure on firms' stock prices, thereby making the firm more likely to be acquired. The correlation between vertical integration and links breaking prevails for vertical acquisitions that follow situations where such fund outflow events put downward pressure on the bottleneck supplier's stock price. If the investor capital outflows are unrelated to the performance of the supplier, these cases are integration events that are unlikely to happen for supply assurance reasons (as, for example, in Bolton and Whinston (1993)). We find similar results when conditioning on situations where the suppliers are "healthy" in the sense that they have seen sales increases prior to integrating.

Moreover, we study events where firms are rumored to vertically integrate or announce an integration, but end up not integrating. To the extent that these rumored integration events might be caused by the same unobserved shocks as actual integration events, they make for a good comparison group. For relationships where suppliers are rumored to vertically integrate, we do not find a higher hazard rate of links breaking than for the average relationship. We also do not find the large difference in hazard rates between rumored integration with a competitor of the buyer versus firms unrelated to the buyer.

To investigate whether strong synergies among merging firms force the downstream competitor out of the market and therefore break the link itself, we study the sales response of firms whose *competitor* is vertically integrating but who did not have a prior relationship with the integrating supplier. We find no statistically significant drop in sales for these firms, suggesting that strong synergies are unlikely to explain the breaking of vertical relationships in our main result. This is consistent with the results of Blonigen and Pierce (2016), who find no significant

increases in physical productivity among US plants that undergo a merger or acquisition, but an increase in market power as measured by markups.

We then use our production network data to ask whether firms that have a foreclosure motive are more likely than others to integrate with a given supplier. We say that a firm b has a foreclosure motive when one of its suppliers also sells to one of b's competitors. In the sample of active relationships where the supplier is vertically integrating, such firms b are more likely to end up being the ones that integrate with the supplier. Again, these results are consistent with foreclosure motives for integration.

Finally, we study the performance of firms in the wake of their supplier's integration. Firms which have a supplier that vertically integrates with one of its competitors experience a temporary decrease in sales. The sales drop is larger for firms that do not have another supplier from the same industry as the one that is integrating. Diversification of the supplier base is hence a possible way to mitigate the impact of being foreclosed.

We interpret our results as supporting the view that vertical market foreclosure along the extensive margin (in the sense that relationships fully break) is occurring in the population of firms and relationships that we study. These relationships are not representative of the overall population of buyer-seller relationships in the United States, or among industrialized countries: the set of firms reporting relationships in our data consists mostly of firms that are either listed on exchanges or issue traded securities. Those firms are also more likely to report relationships with important suppliers and customers. Given that the relationships in our sample will be more likely to be in the spotlight of antitrust authorities, we think that vertical foreclosure may also be prevalent outside the selected sample that we study.

Our paper relates to three different literatures. The first is the literature that studies the determinants and effects of mergers and acquisitions, both domestic (Malmendier et al. (2018), Maksimovic et al. (2013), Rhodes-Kropf and Viswanathan (2004), Gugler et al. (2003), Shenoy (2012), Blonigen and Pierce (2016), Cunningham et al. (2018), Harford et al. (2019)) and international (Blonigen (1997), Nocke and Yeaple (2007), Ekholm et al. (2007), Breinlich (2008), Guadalupe et al. (2012), Stiebale (2016)). In contrast to most of this literature, we study the impact not on integrating firms themselves, but on the vertically related ones.[6] We also show that foreclosure considerations — as determined by the structure of the production network — predict vertical mergers.

The second is the empirical literature on detecting vertical market foreclosure. Waterman and Weiss (1996), Chipty (2001), and Crawford et al. (2018) (in the cable TV industry) and Hastings and Gilbert (2005) (in the gasoline retailing industry) find evidence for vertical foreclosure; Hortaçsu and Syverson (2007) (in cement and ready-mixed concrete markets) and Asker (2016) (in the beer industry) find no vertical foreclosure in their respective industries. In contrast to this literature, we study a range of industries, which not only broadens the scope of statements that we can make, but also allows for comparisons across industries by their

[6]Recent exceptions are Gugler and Szücs (2016) and Stiebale and Szücs (2017), who study the impact of mergers on horizontally related firms.

degree of competitiveness. We draw from data on vertical and competitor relationships, which ties our hands on the definition of markets and vertical integration. The drawback is that our data prevents us from studying prices or markups, and therefore consumer welfare. Instead, we look at the supplier network of potentially foreclosed firms, and how the relationship between integration and links breaking varies with market structure in the upstream market.

Finally, our paper also relates to the growing literature on the importance of firm's position in the production network for its performance and exposure to shocks (Barrot and Sauvagnat (2016), Giroud and Mueller (2017), Bernard et al. (2017), Carvalho et al. (2016), Boehm et al. (2015), Tintelnot et al. (2018), Kikkawa et al. (2018)). Alfaro et al. (2016) study the relationship between prices and vertical integration across many industries. Related to our work, Bernard and Dhingra (2015) find increased integration and foreclosure following the 2012 Free Trade Agreement between Colombia and the United States. Our paper shows how the network matters through the strategic incentives of horizontally related firms, and for how the production network itself is shaped by those incentives. We also introduce a new dataset on buyer-seller connections in the U.S. and abroad and document its properties.

The next section describes the data; Sections 3.3 and 3.4 present the econometric evidence.

3.2 Data

We combine three different datasets for our empirical analysis: a dataset describing supply chain and competitor networks, a dataset of mergers and acquisitions, and data on firm sales and employment. The first dataset is FactSet Revere, a panel of almost 900,000 vertical and horizontal relationships of large US and foreign firms. It describes the supplier, customer, and competitor relationships as well as partnerships of a set of large (mostly publicly listed or security-issuing) firms from the US and abroad (we call these companies the "covered" companies). Each relationship is coded with a relationship type, the identity of the firms, and a start and end date. The data vendor collects this information annually through the covered companies' public filings, investor presentations, websites and corporate actions, and through press releases and news reports. Since the relationship data is the main content of the dataset, its coverage is much broader than supplier data in Compustat or Bloomberg. While the data coverage is specifically geared towards large firms, many small and non-listed firms nevertheless show up in relationships with large firms, hence our overall network is much larger than the set of listed firms. Coverage varies by country; data for covered North American companies is available from 2003 to present; Revere starts to cover publicly listed and security-issuing companies from industrialized and major emerging economies (including Europe and China) from around 2007.[7] To the extent of our knowledge, our paper is the first one in the economics literature to use this dataset, so we show summary statistics in more detail than we otherwise would.

[7] See Appendix 3.A for details on coverage by country and year.

FactSet Revere contains thirteen different types of relationships (see Appendix 3.A.1 for more details). We aggregate these relationship types into two networks: a directed network of buyer-supplier relationships (from supplier and customer relationships, as well as distribution, production, marketing, and licensing relationships) and an undirected network of competitors. Moreover, we annualize the relationship data: A relation of any kind is counted as active in a given calendar year if there is at least one day between start date and end date of the relation that falls into that calendar year. The result is a panel of relations that is identified by source company, target company and year.[8]

Table 3.1: Descriptive statistics for the firm network

	Full Sample			Sample of buyers		
	Mean	SD	Max	Mean	SD	Max
# Customers	2.13	8.28	533	3.37	10.72	533
# Suppliers	2.13	10.20	980	3.85	13.48	980
# Competitors	2.16	7.73	381	3.48	10.13	381
Share of domestic customers	0.49	0.45	1	0.48	0.42	1
Share of domestic suppliers	0.50	0.45	1	0.50	0.45	1
Share of domestic competitors	0.46	0.45	1	0.46	0.43	1
Obs. per firm (years)	3.97	4.30	14	6.24	4.28	14
Log Sales	12.00	2.81	20	12.70	2.62	20
Log Employment	6.27	2.56	15	6.90	2.46	15
Firms	180,192			80,287		

Note: Summary statistics for the number of links in the firm network (2003-2016). The left columns summarize the full set of firms in the database, the right columns only those firms that have at least one supplier in the database. We count relations as domestic when both firms are headquartered in the same country. "Observations per firms" summarizes the coverage length of firms. Sales and employment data come from Compustat, Orbis and FactSet Fundamentals. Note that coverage for sales (employment) is lower: 74,511 (73,613) firms in the full sample and 40,576 (40,389) among buyers.

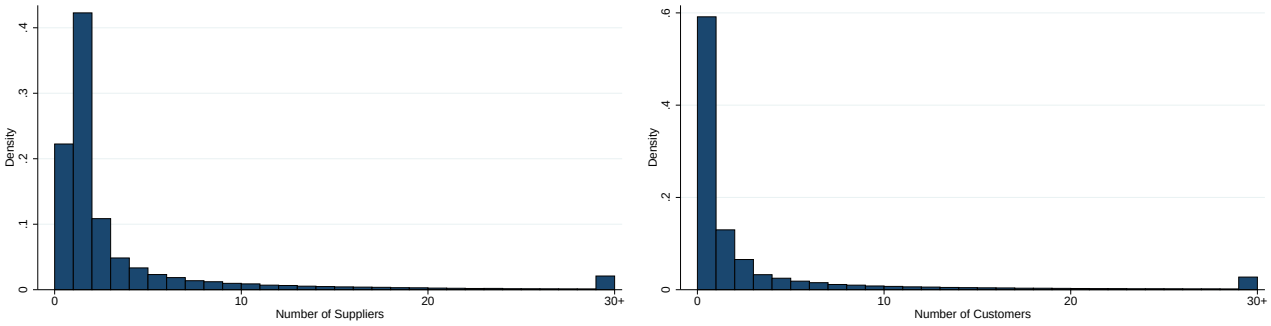
Table 3.1 summarizes the resulting links in the network of firms, which is much more dense than suggested by data exclusively relying on SEC filings (as reported, for instance, by Barrot and Sauvagnat (2016)). Among the more than 180,000 firms in our dataset, 80,000 have at least one supplier link recorded. On average, our buyers have 3.85 suppliers, but many firms have substantially more. The average numbers of customers and competitors is just slightly lower, allowing to construct a dense network. The average length of buyer-supplier relationships in our data is 4.46 years; the unconditional probability of buyer-supplier links breaking in any given year is 22%. Only 6.3% of links that break over the observation period are reformed at a later point in time, and almost never more than once. For buyer-supplier links the share of links that are reestablished later on is higher, at 12.9%.[9]

Figure 3.1 shows the distribution of the number of suppliers and customers among firms.

[8]Firms sometimes undergo organizational changes where a firm identifier ceases to exist and one or more new ones may be created (e.g. in cases of mergers and splits). In such cases, FactSet records the successor identifiers, and we say that a buyer-seller relationship breaks only if there is no buyer-seller relationship with one of the successor firms.

[9]In appendix 3.B.4 we show that our main results are robust to not counting relationships as breaking if they are subsequently reestablished.

Figure 3.1: Distribution of the number of suppliers and customers



Note: The sample consists of firms that have at least one supplier.

The distributions are very skewed, with most firms having few suppliers and customers, and some having many. Whenever we use the number of links in our regressions below, we will hence use the log of one plus the number of links instead of the raw count in order to avoid our results being driven by outliers. The fact that the number of relationships is heavily skewed is well-known from the literatures on firm heterogeneity and superstar firms.[10] Table 3.2 confirms that the firms with most connections account for a disproportionately large fraction of sales.

Table 3.2: Total sales by percentile of the # suppliers distribution

	Fraction of Sales, %
All	100.0
Top 25%	78.1
Top 10%	58.6
Top 5%	46.8
Top 1%	25.8

Note: The table shows the average fraction of sales (over years) accounted for by firms in the top percentiles of the distribution of the number of suppliers (firms with at least one supplier only).

Finally, one word of caution about these data. While the coverage of relationships is better than in other large panels that span many industries and countries, it is probably still incomplete: relationships with small firms, and relationships that account for a small fraction of sales or costs are presumably less likely to be recorded. Our data show about 500 listed suppliers for Walmart in 2016, and Walmart is — together with Apple, Samsung, and the large auto manufacturers — one of the firms with the highest number of recorded suppliers. In reality though, Walmart probably has tens of thousands of suppliers, suggesting that many relationships are missing. The relationships recorded in our data are probably the larger or more important ones.[11]

[10] The literature is vast; see, in particular, the recent empirical work by Bernard et al. (2017). Most similar to us, Carballo et al. (2018) document the skewness of the customer distribution and sales for international buyers of Latin American firms. In theoretical work, Oberfield (2018) explains how superstar firms emerge in a setting where firms search for suppliers.

[11] Alternatively, one could use administrative VAT transaction records, as are available for countries like Belgium (Bernard et al. (2017)) and Chile (Huneeus (2018)). However, in that case our study would be limited

The second dataset we use is the set of mergers and acquisitions in Bureau Van Dijk's Zephyr database. Zephyr records deals and rumors about deals for mergers and acquisitions in which at least a 2% stake in the target company changes owners and the deal's value exceeds GBP 1M (Bollaert and Delanghe (2015)). For each merger or acquisition, Zephyr reports the nature of the transaction, the identity of the target company, the acquiring company and the seller, as well as the date of announcement, the date when the transaction was finished, and the stake of the acquirer in the target before and after the acquisition. Zephyr also contains a large number of rumored deals that never materialized, which we will use as a comparison group in some of our regressions.

Analogously to the relationship data, we annualize the Zephyr data and construct a panel of mergers and acquisitions between acquiring and acquired company. We focus on transactions where one company fully acquires another or the entities merged. We infer the vertical or horizontal nature of an integration by combining the M&A data with the input-output network: a vertical integration is a merger or acquisition between two firms that have an ongoing buyer-seller relationship in the year of integration.

The vast majority of mergers and acquisitions in our sample is between firms that do not maintain a buyer-supplier relationship. Table 3.3 reports the number of mergers and acquisitions between firms for which supply chain information is available. Only 6.7% of full acquisitions in our sample result in vertical integration. The share is almost the same for partial acquisitions, which we do not use in our analysis but report here for completeness. The non-vertical mergers and acquisitions can be either purely horizontal or between unrelated firms that neither compete directly nor supply each other with inputs. For the sake of brevity, we will refer to both mergers and acquisitions as "mergers" for the remainder of the paper.

Table 3.3: Types of mergers and acquisitions

	Non-vertical		Vertical		Total	
	Count	%	Count	%	Count	%
Partial acquisitions	745	93.2	54	6.8	799	100.0
Full acquisitions & mergers	2,799	93.3	201	6.7	3,000	100.0
Total	3,544	93.3	255	6.7	3,799	100.0

Note: Number of partial and full mergers and acquisitions by presence vertical relation between the merging parties (2003-2016). Partial acquisitions exclude minority stakes. For a breakdown including horizontal mergers see Appendix 3.A.

There is a small but non-negligible number of cases with risk of vertical foreclosure. Table 3.4 summarizes key statistics about the buyer-supplier relations in our sample. While the unconditional probability that a relation ends in a given year is only 22.4%, this probability is more than 50% in cases where the supplier integrates vertically with a competitor of the buyer. In our data, this happens in 102 out of the 6613 cases in which a supplier vertically integrates with another buyer.

to relatively small samples and few vertical merger cases (as well as additional constraints imposed by the

Table 3.4: Buyer-supplier links: hazard rates of links breaking and risk of foreclosure

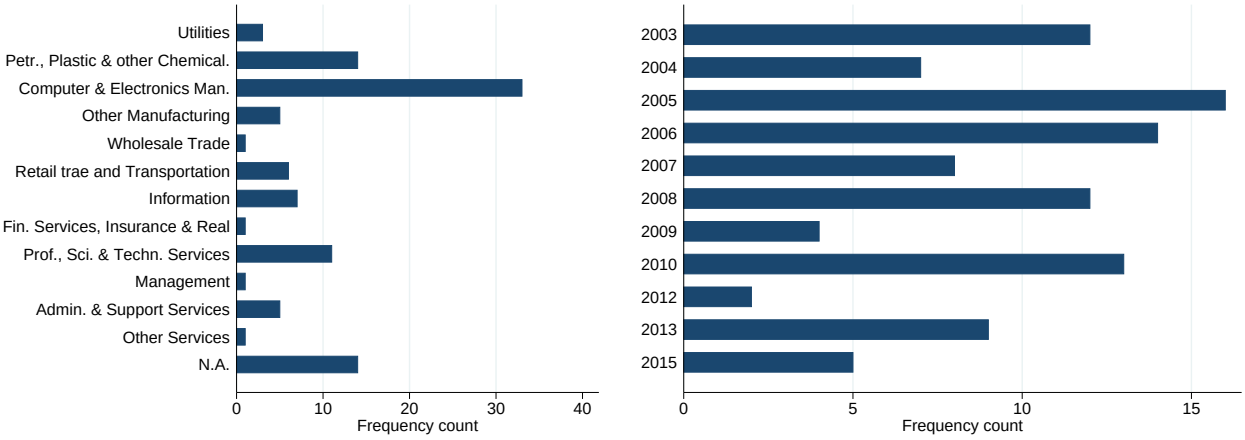
	Value
P(link breaks)	0.224
Avg. relatation duration	4.46
Number of cases where supplier vertically integrates	6613
Number of cases where supplier integrates w. competitor	199
... and buyer-supplier link breaks	102

Note: The first row reports the unconditional probability that a buyer-supplier relationship ends in a given year. The second row reports the average length of these relations. The third row counts the number of cases in which a supplier vertically integrates. The fourth row restricts this number to cases where the vertical integration involves a competitor of the buyer. The fifth row counts the instances in which the buyer-supplier link breaks following vertical integration of the supplier with a competitor of the buyer.

Figure 3.2 shows the industry-wise and year-wise distribution of cases where the relationship breaks following vertical integration of the supplier with a competitor of the buyer. These situations are not confined to a narrow set of industries, but occur broadly across the economy. A particularly large number of such cases falls into computer and electronics manufacturing, in which there are many large firms that are frequently undertaking mergers and acquisitions.

In the short panel that is available to us, there is no clear trend over time in the number of potential foreclosure cases. Whereas recent research has documented a rise market power since the early eighties (De Loecker et al. (2018)), this does not translate into an increase in the number of potential foreclosure cases over time in our sample.

Figure 3.2: Potential foreclosure cases by sector and year



Note: A potential foreclosure case is a situation where a buyer-seller relationship breaks following integration of the supplier with a competitor of the buyer. About three quarters of potentially foreclosed firms are US firms.

We complement the relationship and M&A data by sales and employment figures and industry codes from Compustat, Bureau Van Dijk's Orbis database and FactSet Fundamentals (2003–2014). Since these data have been widely used in the literature, we will not describe

confidential nature of these data).

them here.[12] The last rows of Table 3.1 show summary statistics for sales and employment.

3.3 Extensive-margin foreclosure

3.3.1 Empirical strategy

Our empirical strategy is to study whether vertical relationships are more likely to break after the supplier integrated with a competitor of a buyer, than when it integrated with an unrelated firm. Consider a vertical relationship between seller s and buyer b. If b is a competitor of the firm b' that s is integrating with, then the integrating parties may have an incentive to foreclose b. If, on the other hand, b and b' are in different markets, then b would not be threatened by foreclosure (see Figure 3.3). Our strategy is therefore to compare the probability of the (b, s) relationship breaking between these two scenarios.

We define markets through the competitor relationships that we observe in FactSet Revere. FactSet constructs these competitor relationships based on firm's product portfolios and self-disclosed competitor relationships from SEC filings. We prefer this definition over industry code-based definitions for two reasons. Firstly, even 6-digit NAICS categories are often broad and encompass many different product markets (e.g. NAICS 334310: "Audio and Video Equipment Manufacturing"). Secondly, many of the firms in our sample are large firms that operate in different product markets, which are not always reflected in the SIC or NAICS codes. For those firms, competitor relationships are usually not transitive. As a result, the FactSet competitor relationships are very different from co-memberships in industry cells: among competitor pairs according to FactSet, only 43.5% are among firms that share a 4-digit NAICS code. Conversely, among all pairs of firms that share a 4-digit NAICS code, only 0.03% coincide with a FactSet competitor link.[13]

We estimate the following linear probability model[14] on the set of all triples (b, s, t) where s is listed as one of b's suppliers (or b is listed as one of s's customers) for at least one day in year t:

$$\begin{aligned}\mathbb{1}\{\text{LinkBreaks}\}_{bst} = \; & \alpha \mathbb{1}\{s \text{ vertically integrates}\}_{st} \\ & + \beta \mathbb{1}\{s \text{ integrates vertically w. competitor of } b\}_{bst} \\ & + \eta_{bs} + \eta_{bt} + \eta_{i(b)i(s)t} + \varepsilon_{bst} \end{aligned} \tag{3.1}$$

where $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$ is a dummy variable that is one if and only if the vertical relationship

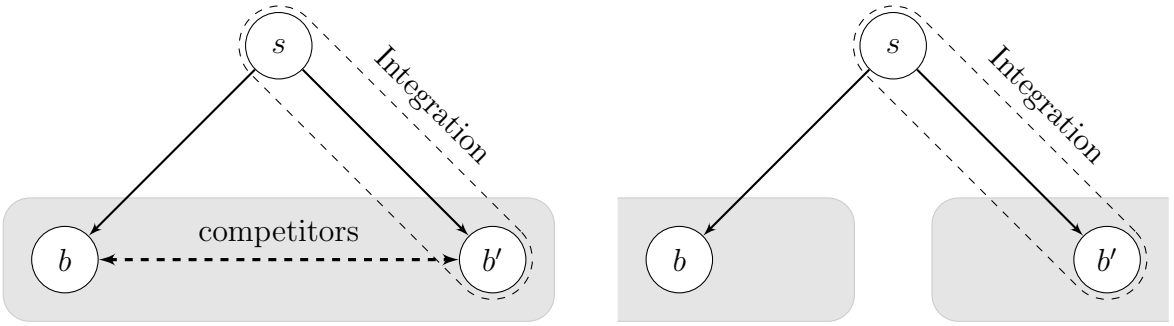
[12] See Kalemli-Ozcan et al. (2015) for detailed information on Orbis. We use a current and past vintage of Orbis to have a better coverage.

[13] The corporate finance literature is well aware that industry co-membership is a poor way to measure competitor relationships. Rauh and Sufi (2011) use competitor definitions from CapitalIQ and argue that this method captures competitor relationships much more accurately than using industry codes. Hoberg and Phillips (2010) develop measures of product market competition from text analysis of firm filings.

[14] We use a linear model as a benchmark specification because it allows us to include high-dimensional fixed effects. We estimate hazard models in Appendix 3.B.3, and find similar results.

between b and s is active during year t, but not during year $t+1$ (and also not between entities that are successors to b or s in case of a split or change in organizational form). The right-hand side variables are a dummy for whether s vertically integrates during year t, and a dummy for whether s vertically integrates with a competitor of b during year t. We include (i) fixed effects for buyer × year, η_{bt}, to control for time-varying characteristics of the buyer that could make all its supplier relationships more likely to break during a given year (such as exit), (ii) buyer × supplier fixed effects, η_{bs}, thereby identifying the coefficients of interest, α and β, from within-relationship variation in the hazard rate of the relationships breaking, and in the firms' characteristics, and (iii) industry-pair × year fixed effects, $\eta_{i(b)i(s)t}$, which takes out industry-specific (or industry-pair-specific) shocks that may lead to a higher break probability (where industries are defined at the 3-digit NAICS level). We exclude relations from the regression where the buyer and supplier themselves are the vertically integrating parties.

Figure 3.3: Empirical strategy: compare situations where buyers b and b' are in same vs. different markets



(a) Buyers b and b' in same market: foreclosure potential

(b) Buyers b, b' in different markets: no foreclosure potential

Note: This figure illustrates the main empirical strategy. We compare two situations in which a seller s integrates with a buyer b': one in which b and b' compete in the same product markets (a) and one in which they do not (b).

Table 3.5 shows the result from estimating equation 3.1 using ordinary least squares. The first column shows that when suppliers are vertically integrating, the probability of a given vertical relationship breaking is higher by about 2.1 percentage points (though this is not statistically significant). Given that the unconditional probability of a relationship breaking in our data is about 22%, this would constitute an increase of about 9%. Column (2) shows that the likelihood of the vertical relationship breaking is indeed much higher (18 percentage points difference, or a 80% higher probability) when the buyer is a competitor to the downstream merging firm. This difference remains large and statistically significant when including industry pair × year fixed effects to control for sector- or sector-pair-specific shocks (column 3), and when controlling for a range of supplier and relationship characteristics (column 4).

It is worth pointing out that the results above are unlikely to be driven by the possibility that a relationship may not be observed by FactSet following a merger, because the firm entity has ceased to exist, or because it may not be tracked anymore: if that was the case, we should

Table 3.5: Correlation of buyer-supplier link breaking with vertical integration of supplier

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$			
	(1)	(2)	(3)	(4)
Supplier v. integrates	0.021	0.013	0.005	0.020
	(0.019)	(0.019)	(0.020)	(0.018)
Supplier v. integrates w. competitor		0.181**	0.178**	0.148**
		(0.059)	(0.062)	(0.051)
Controls				Yes
Relation FE	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes
Industry Pair × Year FE			Yes	Yes
R^2	0.578	0.578	0.619	0.671
Observations	640725	640708	472763	472763

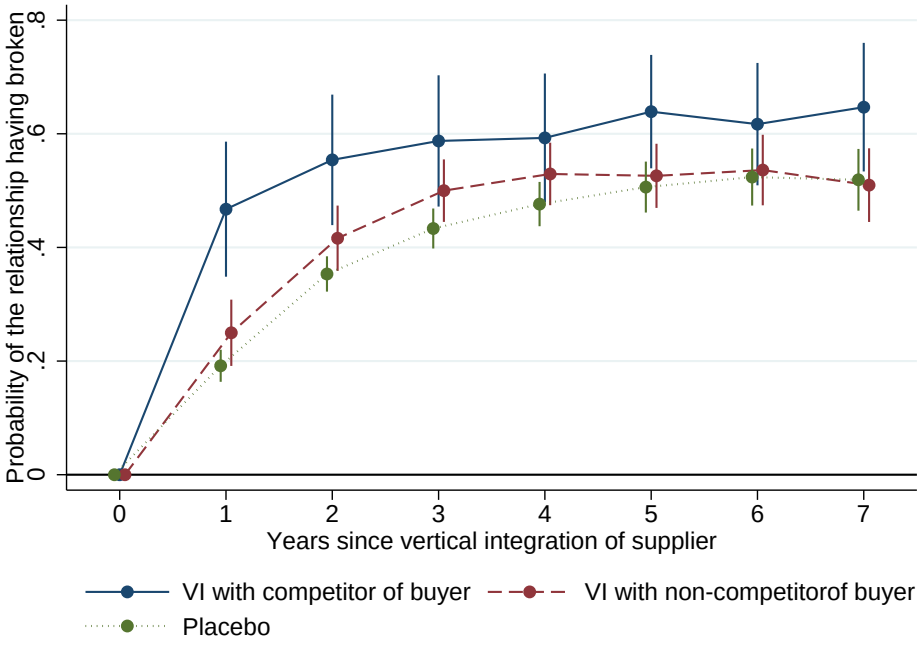
Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the supplier-year level. The number of reported observations is the number of non-singleton observations. The drop in the number of observations in columns (3) and (4) is explained by firms with missing industry codes.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

be seeing a substantially increased hazard also following vertical mergers with firms that are not competitors of the downstream merging firm.

Figure 3.4 shows graphically how break probabilities differ across these two types of vertical integration events. The horizontal axis shows the time after a vertical integration event of the supplier; the vertical axis shows the probability of the relationship having broken (i.e. one minus the probability of the relationship being active). By definition of the sample, in the year of integration of the supplier the buyer-seller relationship must be active. We see that relationships where the supplier integrates with a competitor of the buyer (solid blue line) are much less likely to survive the post-integration years, in particular the year following integration, than relationships where the supplier integrates with a non-competitor of the buyer (dashed red line). The dotted green line shows relationship survival rates for simulated placebo events that are generated to occur with 0.6% probability in any given year where a relationship is active. This corresponds to the average probability that a supplier in a given relation vertically integrates. The regression that generates these marginal effects include relationship, buyer-year, and industry-pair year fixed effects; the corresponding plot of a regression without fixed effects looks very similar.

Next, we study variation across industries in the relationship between vertical integration and links breaking. Most theories of vertical foreclosure, in particular the raising rivals' cost theories and extending monopoly power theories of vertical foreclosure predict that market power in the bottleneck market increases the incentives to foreclose. We want to empirically assess this prediction. In order to do so, we study whether the correlation between integration with a competitor and relationships breaking is lower when the supplier has less market power.

Figure 3.4: Probability of relationships having broken after supplier's vertical integration



Note: The figure shows coefficients on dummies capturing the years since a supplier's vertical integration, in a regression of the probability of a buyer-seller relationship being inactive on time-since-integration dummies, as well as relationship, buyer × year, and industry-pair × year fixed effects. Standard errors are clustered at the supplier-year level. The solid blue line denotes relationships where the supplier integrates with a competitor of the buyer; the dashed red line denotes relationships where the supplier integrates with a non-competitor of the buyer; the dotted green line represents relationships where a placebo integration event has been drawn to occur. That placebo event is randomly drawn to occur with 0.6% probability in any given year where a relationship is active (and independently across relationship-years).

We measure the supplier's market power by the number of his competitors.[15] More specifically, we run the regression

$$\begin{aligned}\mathbb{1}\{\text{LinkBreaks}\}_{bst} = \; & \alpha \mathbb{1}\{s \text{ vertically integrates}\}_{st} \\ & + \beta \mathbb{1}\{s \text{ integrates vertically w. competitor of } b\}_{bst} \\ & + \gamma \mathbb{1}\{s \text{ integrates vertically w. competitor of } b\}_{bst} \times C_{st} \\ & + \delta C_{st} \\ & + \eta_{bs} + \mu_{bt} + \varepsilon_{bst} \end{aligned} \tag{3.2}$$

where C_{st} is a variable capturing the number of competitors of the supplier s at time t. Just like the number of buyers and suppliers is heavily skewed, so is the number of competitors, therefore we use the log of one plus the number of competitors for C_{st}.

Table 3.6 shows the results. We find that the correlation between buyer-supplier-links

[15] Alternatively, one could measure the supplier's market power with market shares. Our sales coverage among suppliers and in upstream markets generally, however, is very limited, so we prefer measuring supplier market power through the number of competitors.

breaking and vertical integration of a supplier with a competitor is lower when the supplier has more competitors (columns (1) and (2)). This result is in line with theories of foreclosure: the existence of more alternative suppliers to the buyer reduces the incentives of the acquirer to foreclose competitors. In columns (3) and (4) we also include interactions with the number of competitors of the buyer. Perhaps surprisingly, the point estimates of the coefficients on these interaction terms are slightly positive (though not statistically significant). While not being entirely conclusive, it does not seem to be the case that more competition in the downstream market reduces the probability of links breaking after integration with a competitor. This stands in contrast to theories where foreclosure arises to preserve market power on the downstream market.

Table 3.6: Interaction with the number of upstream competitors

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$			
	(1)	(2)	(3)	(4)
Supplier v. integrates w. competitor	0.562**	0.461**	0.325	0.246
	(0.171)	(0.160)	(0.342)	(0.312)
Supp. v. int. w. comp. × # upstream comp.	-0.127*	-0.111*	-0.126	-0.110
	(0.057)	(0.054)	(0.093)	(0.086)
Supplier v. integrates	0.008	0.025**	0.008	0.025
	(0.008)	(0.007)	(0.020)	(0.022)
# upstream competitors	-0.016**	-0.023**	-0.016**	-0.023**
	(0.002)	(0.002)	(0.003)	(0.003)
Supp. v. int. w. competitor × # downstream competitors			0.064	0.058
			(0.047)	(0.040)
Controls		Yes		Yes
Relation FE	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes
Industry Pair × Year FE	Yes	Yes	Yes	Yes
R^2	0.619	0.667	0.619	0.667
Observations	472763	472763	472763	472763

Note: Controls: number of upstream customers, age of the link, dummy indicating other links of the supplier breaking. "Upstream competitors" is the number of competitors of the supplier; "downstream competitors" is the number of competitors of the buyer. Table reports robust standard errors, clustered at the supplier-year level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Tables 3.5 and 3.6 show a correlation that by itself is not evidence for vertical foreclosure. We see that relationships are relatively much more likely to break when the supplier is undergoing a vertical merger with a competitor of the buyer, than when it is merging with a firm that is not competitor of the buyer. The fact that this correlation is stronger when the supplier has few competitors lends support to the view that vertical foreclosure along the extensive margin could be occurring in the population of firms that we study. Yet, the regressions are not necessarily evidence for a causal link between mergers and the breaking of relationships, simply because mergers do not happen randomly. In particular, there are three main confounding explanations:

Firstly, it could be that the integration between the supplier and the competitor is a consequence of the relationship between buyer and supplier breaking; for instance because the supplier's acquirer might be concerned that the supplier would otherwise exit.[16] In that case our regression would suffer from reverse causality: integration with a competitor of the buyer would be relatively more likely because the competitor could be purchasing exactly those goods that the supplier is discontinuing.

Secondly, it could be that both the breaking of the relationship and the vertical integration are the result of an unobserved shock hitting one of the firms. Such a shock would need to make the supplier more likely to integrate with competitors of its buyers than with a non-competitor in order to explain the different magnitude of the coefficient estimates in Table 3.5. We discuss these alternative explanations in turn.

Thirdly, if synergies between the vertically integrating firms are very strong, the resulting cost savings in the production of the downstream good could drive their competitors in the downstream market out of the product market, and lead them to stop buying from the upstream integrating firm.

We proceed to discuss the first two alternative explanations, and turn to the third one after showing the impact of separations on sales in Section 3.4.

3.3.2 Reverse causality: vertical integration for supply assurance?

Our relationship between links breaking and vertical integration may be driven by suppliers' motivation to exit certain product markets and cut ties with some of their customers, which in turn may cause them to be acquired by one of their customers. We therefore apply an instrumental variable strategy that exploits shocks that are outside of the control of the firm and that make integration more likely. Our instrument builds on Edmans et al. (2012), who show that when large mututal funds experience an outflow of capital, they are forced to sell off assets, which puts downward pressure on the share prices of firms in their portfolio. In turn, these firms become more likely to be acquired.

We follow Edmans et al. (2012) and Dessaint et al. (2016) and construct a variable capturing the *hypothetical* (not actual) share sales of large U.S. mutual funds in response to an outflow of investor capital. We first calculate the net inflow of capital to the fund based on its total net asset holdings and returns reported in the CRSP mututal funds database. For funds j that see a net outflow of more than five percent of its total net assets in a given quarter q, we calculate the hypothetical sales of a stock i if holdings of all assets were reduced proportionally to the

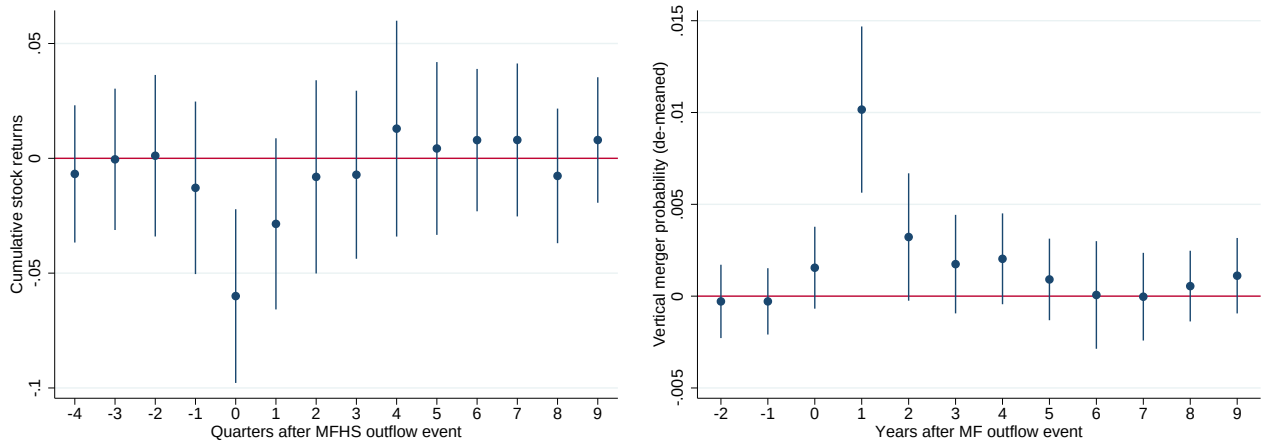
[16] Bolton and Whinston (1993) study firms' incentives to vertically integrate for supply assurance reasons. In this situation, "exit" does not have to be a complete exit of the supplier, but could be just an exit from a particular market.

outflow.[17] The total hypothetical sales of a stock i from mutual fund outflows are then

$$MFHS_{i,q} = \sum_{j:\ \text{Flow}_{j,q}<-0.05} (\text{Flow}_{j,q} \cdot \text{Shares}_{ji,q-1} \cdot \text{Price}_{i,q-1})$$

We sum this variable over the four quarters in the year and normalize the sum by the total trading volume in that year.

The normalized MFHS variable is meant to capture the downward pressure on prices that is exerted by the fund's capital outflow. Figure 3.5a shows the average response of cumulative stock returns following a large mututal fund outflow event (defined as normalized MFHS below the tenth percentile). Stock prices drop significantly as the shock hits and then recover to the pre-shock level. Figure 3.5b shows the response of the probability to be involved in the completion of a vertical merger or acquisition before and after such an event. In the year after the outflow event, the probability of integration is significantly higher. The one year lag between outflow event and completion of the acquisition may reflect the time to negotiate the acquisition and the antitrust authority's clearance.



(a) Cumulative stock returns

(b) Vertical merger probability

Figure 3.5: Response to a mutual fund outflow event

Note: The figures show the average response of cumulative stock returns (vertical axis, left panel), and the average response of the probability to engage in a vertical merger or acquisition (vertical axis, right panel) following a mutual fund capital outflow (defined as normalized MFHS being below the tenth percentile) at quarter 0. Both regressions contain firm and industry-time fixed effects; standard errors are clustered at the firm level.

Table 3.7 shows the results of estimating equation (3.1) with the interaction terms instrumented by an interaction of the competitor status with a dummy that is one if the vertical integration happens up to two years after a mutual fund outflow event (which, as the construction suggests, happens disproportionately often: in about a third of our cases of integration with a competitor). This instrument effectively limits the set of vertical mergers that are being considered to post-outflow vertical mergers, which are much less likely to be driven by the per-

[17] Data on mututal fund stock holdings come from the Thomson Spectrum CDA database, and stock prices from Thomson Worldscope. See Appendix 3.A for data sources and definitions.

formance of suppliers or buyers. The estimated coefficient on the variable representing vertical integration with a competitor of the buyer remains large and statistically significant, suggesting that our baseline results are not driven by the possibility that integration is the response to links breaking. While the point estimates are slightly larger than in our baseline, they are also less precise. It is therefore not clear whether OLS was biased in the first place. Indeed, an overidentification C-test fails to reject the null hypothesis that the regressors are exogenous ($p = 0.11$).

Table 3.7: Relationships breaking following Vertical Integration: IV results

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$		
	(1)	(2)	(3)
Supplier v. integrates	-0.002	-0.008	0.006
	(0.021)	(0.021)	(0.021)
Supplier v. integrates w. competitor		0.262***	0.202*
		(0.074)	(0.081)
Controls			Yes
Method	IV	IV	IV
Relation FE	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes
Industry Pair × Year FE			Yes
R^2	0.578	0.578	0.671
Observations	640725	640708	472763

Note: This table shows regressions where the interactions are instrumented by an interaction of the competitor dummy with a dummy that is one if the vertical merger happens up to and including two periods after a mututal funds outflow event. This effectively reduces the explanatory variable to include only post-outflow vertical mergers (instead of all vertical mergers). Because of this interaction, the first-stage Kleibergen-Paap F-statistic is large. Robust standard errors, clustered at the supplier-year level, are in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

As an alternative to the IV strategy, we show results where we restrict attention to a subsample of firms that are "healthy", and are therefore less likely than the average firm to cut substantial parts of their product mix.

Table 3.8 shows results of estimating equation (3.1) on the sample of firms that have positive sales growth between years $t-2$ and $t-1$ (columns (1) to (3)), or sales growth above the median of three percent (columns (4) to (6)). The point estimates of the coefficient on the integration with a competitor variable are larger than in our baseline specifications (even though the smaller sample makes the estimate less precise). Firms that are growing are much less likely to exit product markets (Goldberg et al. (2010)). For the firms in this subsample, the causality is hence much less likely to run from the breaking of the relationship to vertical integration.

3.3.3 Unobserved shocks: omitted variables

Our next exercise speaks to the possibility that both vertical integration and the discontinuation of buyer-supplier relationships are the response to unobserved shocks. As discussed above, such

Table 3.8: Regressions on relationships with "healthy" suppliers

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$					
	Sample: $\Delta \log \text{Sales}^s_{t-1} > 0$			Sample: $\Delta \log \text{Sales}^s_{t-1} > \text{median}$		
	(1)	(2)	(3)	(4)	(5)	(6)
Supplier v. integrates	0.033	0.021	0.033	0.048	0.022	0.036
	(0.026)	(0.028)	(0.025)	(0.032)	(0.035)	(0.030)
Supplier v. integrates w. competitor	0.387**	0.313*	0.213+	0.373**	0.361*	0.238
	(0.118)	(0.129)	(0.122)	(0.144)	(0.155)	(0.146)
Controls			Yes			Yes
Relation FE	Yes	Yes	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry Pair × Year FE		Yes	Yes		Yes	Yes
R^2	0.606	0.674	0.709	0.616	0.685	0.719
Observations	251966	191654	191654	197738	148121	148121

Note: Columns (1) to (3) restrict the sample to buyer-supplier pairs where $\Delta \log \text{Sales}^s_{t-1}$ is above zero, columns (4) to (6) where it is above the median. Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Number of observations exclude singleton observations. Robust standard errors, clustered at the supplier-year level, in parentheses.
$^+$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

shocks must be directed to make integration with a competitor of the buyer more likely in order to explain the correlation in the baseline tables. One could think of one buyer making an innovation which increases the need for customization of the supplied input, while also driving the competitor out of business. The innovator and supplier choose to vertically integrate to reduce the inefficiency associated with the hold-up problem (Klein et al. (1978)).

We try to find a group of firms that is most comparable in terms of the shocks that they may have been facing, but for an *exogenous* reason do not manage to vertically integrate. The closest we can get to such a comparison group is by considering rumors of mergers and mergers that have been announced, but for some reason have not been completed. Zephyr collects the former from "unconfirmed reports", which "may be in the press, in a company press release, or elsewhere" (Bureau Van Dijk (2017)). Our approach is hence similar to the comparison of a placebo with the actual treatment in the sense that our rumor or attempted merger does not actually result in vertical integration (but potentially with the difference that even an attempted merger may lead to buyers switching suppliers). Rumors are dated at the time when they are first mentioned. While buyers in rumored and actual treatments are quite comparable, the suppliers that are rumored to integrate are somewhat larger than the suppliers that actually integrate (see Table 3.18 in the appendix). Note that we can control for these differences in our regressions and also do not find differential effects for larger or smaller suppliers.

We first study the benchmark specification, equation (3.1), with actual vertical integration events replaced by the rumors and announced but not completed mergers.[18] This specification compares the average probability of links breaking outside of such events with the average break

[18] We do not count a merger as a rumor if it has been later announced and completed.

probability under a rumored vertical integration, and one with a competitor of the buyer. Table 3.9 reports the results of these regressions. Links break slightly less often during rumored vertical integration with non-competitors of the buyer, and slightly more often (though not statistically significantly so) during rumored vertical integration with competitors. The point estimate of the coefficient on the "rumored vertical integration with competitor" dummy is certainly much lower that the corresponding point estimate in the benchmark regression with actual mergers (though note that the comparison is not straightforward: the dummy here is one at the rumor or announcement date, whereas it is one in Table 3.5 on the *completion* date). Table 3.21 in Appendix 3.B.2 shows results with both rumors and actual integration events in the same regression.

Table 3.9: Links are not more likely to break following rumors of M&A

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$	
	(1)	(2)
Supplier v. integrates (rumor)	-0.030^{+}	-0.029^{+}
	(0.018)	(0.016)
Supplier v. integrates w. competitor (rumor)	0.031	0.020
	(0.039)	(0.035)
Controls		Yes
Relation FE	Yes	Yes
Buyer × Year FE	Yes	Yes
Industry Pair × Year FE	Yes	Yes
R^2	0.586	0.639
Observations	596656	596657

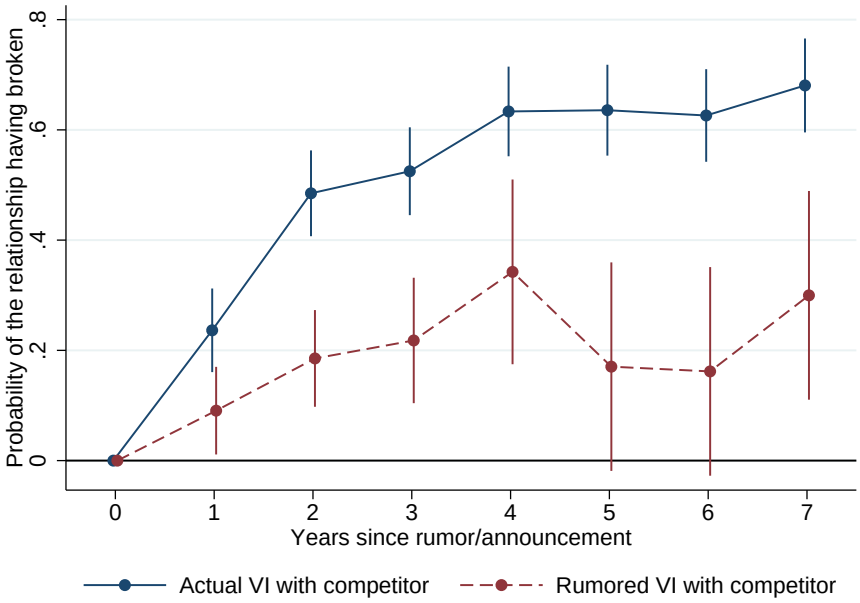
Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Number of observations exclude singleton observations.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

To investigate more closely the timing aspect and to have the tightest possible comparison between actual and rumored/attempted mergers, we compare the break probability before and after actual mergers with buyers' competitors to the break probability before and after rumored/attempted mergers with buyers' competitors. In both cases we use the date of the announcement. More precisely, we run a regression of a binary variable that is one if the relationship is not active anymore on a set of dummies for the number of years since announcement, separately for actual and rumored mergers (and separately by whether the merger is with a competitor of the buyer), and including relationship, buyer × year, and sector-pair × year fixed effects.

Figure 3.6 shows the results. Following the announcement, break probabilities are substantially higher for actual than for rumored vertical mergers with competitors. Not only are relationships where there is a rumor about the supplier integrating with a competitor not more likely to break in the first period, but these relationships seem to be fairly long-lasting. To the extent that rumors and situations in which announced mergers are unsuccessful are a good

comparison group to actual merger events, vertical integration and links breaking are unlikely to be driven by the same underlying unobserved shocks.

Figure 3.6: Probability of relationships breaking: actual vs rumored integration with competitor



Note: The figure shows coefficients on dummies capturing the years since a supplier's rumored (dashed red line) or actual (solid blue line) vertical integration, in a regression of the probability of a buyer-seller relationship being inactive on time-since-integration dummies (separately for rumored mergers with competitors, with non-competitors, and actual mergers with competitors, and with non-competitors) as well as relationship, buyer × year, and industry-pair × year fixed effects. Here, time zero is the time of the rumor or the announcement of the merger. We exclude rumors that are realized within three years.

3.3.4 Is foreclosure a merger motive?

The correlations presented above are consistent with theories of vertical market foreclosure. That said, even if the timing of a vertical integration of the supplier is exogenous, the party with whom the supplier integrates may not be unrelated to firm or market structure: an acquirer that senses a foreclosure opportunity may be willing to pay a premium, and is therefore more likely than alternative bidders to be the winning bidder.

To study whether vertical foreclosure is a merger motive, we run the regression

$$\begin{aligned}\mathbb{1}\{b \text{ integrates with } s\}_{bst} = \ & \alpha \mathbb{1}\{b \text{ has a competitor that is supplied by } s\}_{st} \\ & + \eta_{st} + \varepsilon_{bst}\end{aligned} \tag{3.3}$$

on the sample of active buyer-supplier relationships (b, s) at time t when the supplier s is undergoing a vertical integration with one of its customers. The coefficient α tells us whether buyers that have a competitor that is also a customer of the supplier are more likely to be the one that is integrating with the supplier — conditional on the supplier vertically integrating. These buyers potentially have a motive to foreclose their competitors.

Table 3.10 shows the results. The point estimate of α is positive and statistically significant. Given that the unconditional probability of being the integrating party in this sample is about three percent, having a foreclosure motive is associated with a roughly 55% higher probability of being the firm that integrates with the supplier. In column (2) we control for the buyer's (log of one plus the) number of suppliers and competitors, which proxies for size and alleviates the concern that buyers with a competitor among the seller's customers are just those that are larger. In column (3) we include dummies for the buyer's industry times year, to control for industry-time-specific shocks. Neither of these controls affect the estimate of α much. Hence, firms that have a foreclosure motive (in the sense that s is also supplying their competitor) are more likely to be the integrating party at a time when s vertically integrates.

Table 3.10: Buyers with competitors that are also supplied by S are more likely to integrate with S

	Dependent variable: $\mathbb{1}\{\text{B and S integrate}\}_{bst}$		
	(1)	(2)	(3)
B has competitor supplied by S	0.017***	0.017**	0.017**
	(0.004)	(0.005)	(0.006)
Controls		Yes	Yes
Supplier × Year FE	Yes	Yes	Yes
Buyer Industry × Year FE			Yes
R^2	0.101	0.105	0.167
Observations	6812	6812	5960

Note: Sample consists of all active buyer-seller relationships at a time where the supplier vertically integrates with a buyer. Controls: number of buyer's competitors and suppliers. Reported number of observations is net of singleton observations. The drop in the number of observations in column (3) is explained by firms with missing industry codes.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.3.5 International relationships and cross-border mergers

We now turn to studying the international dimension in our regressions. Buyers with foreign suppliers may be at higher risk of foreclosure if competition authorities do not take foreign markets into account in their merger evaluation. Similarly, cross-border mergers, which account for about 20% of full vertical mergers (Table 3.11), may receive a different degree of scrutiny than purely domestic mergers. We therefore look at whether the extensive margin of (cross-border or domestic) relationships correlates differently with integration.

Table 3.12 shows the results. The first two columns are the same as in Table 3.5 with the difference that we add country-pair × year fixed effects. Columns (3) and (4) include interactions with a dummy that is one if b and s are registered in different countries. Whereas the coefficient on the interaction with any kind of vertical integration by a supplier is negative and weakly statistically significant, we do not find evidence suggesting that international relations are more likely to become targets of foreclosure. Columns (5) and (6) include interactions with a dummy that is one if the buyer that s is integrating with is located in a different country.

Table 3.11: Domestic and cross-border mergers and acquisitions

	Non-vertical		Vertical		Total	
	Count	%	Count	%	Count	%
Domestic	2,038	92.7	161	7.3	2,199	100.0
Cross-border	761	95.0	40	5.0	801	100.0
Total	2,799	93.3	201	6.7	3,000	100.0

Note: Number of full mergers and acquisitions by presence of a vertical relation between the merging parties (2003-2016). M&As are counted as domestic if both merging parties are headquartered in the same country, otherwise they are considered cross-border M&A.

Their coefficients, too, are small and insignificant. International mergers seem to be no different to domestic mergers when it comes to their likelihood of foreclosing the competition.

3.4 Impact on foreclosed firms

3.4.1 Impact on sales

The results from the previous section show that buyer-seller relationships are more likely to break when the downstream merging firm is a competitor of the buyer. The obvious next question is: does it matter? If the input market is frictionless and perfectly competitive, the cost to losing a supplier is zero (of course, in such a situation there is no foreclosure motive at all). If, on the other hand, the use of outside suppliers is associated with a higher variable cost, then the loss of the supplier will push the buyer along the demand curve to a point where the firm operates at a lower scale.

We now study the response of firm sales to events where (1) a supplier of the firm vertically integrates; (2) a supplier of the firm vertically integrates with a competitor of the firm. Specifically, we estimate the equation

$$\begin{aligned} \log Sales_{bt} = \; & \alpha \mathbb{1}\{\text{A supplier vertically integrates}\}_{bt} \\ & + \beta \mathbb{1}\{\text{A supplier integrates vertically w. competitor of } b\}_{bt} \\ & + \eta_b + \mu_{i(b)t} + \varepsilon_{bt} \end{aligned} \tag{3.4}$$

where η_b is a buyer fixed effect, and $\mu_{i(b)t}$ is an industry $\times$ year fixed effect. While our sales variable is constructed from accounting data and is probably measured with error, this should not bias our estimates as long as the measurement error is classical.

The first two columns of Table 3.13 show the results. In a year where a supplier of the firm is integrating with a non-competitor, the firm's sales are slightly higher; if the integration happens with a competitor, the sales are slightly lower than average. But this small coefficient is masking a lot of heterogeneity. Columns (3) and (4) interact the dummy for vertical integration with a competitor with a variable capturing the number of other suppliers from the same 3-

Table 3.12: International Relationships, International M&A's

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$					
	(1)	(2)	(3)	(4)	(5)	(6)
Supplier v. integrates	0.003	0.018	0.014	0.028	-0.001	0.015
	(0.020)	(0.019)	(0.021)	(0.020)	(0.021)	(0.019)
Supplier v. integrates w. competitor	0.187**	0.158**	0.162*	0.147**	0.182**	0.154**
	(0.063)	(0.052)	(0.065)	(0.054)	(0.065)	(0.054)
S integrates × Intl. Rel.			-0.046*	-0.042+		
			(0.023)	(0.023)		
S integrates w. comp. × Intl. Rel.			0.106	0.045		
			(0.112)	(0.106)		
S integrates × Intl. M&A.					0.078	0.057
					(0.088)	(0.082)
S integrates w. comp. × Intl. M&A.					0.004	0.001
					(0.231)	(0.196)
Controls		Yes		Yes		Yes
Relation FE	Yes	Yes	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry Pair × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Country Pair × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.636	0.683	0.636	0.683	0.636	0.674
Observations	464643	464643	464643	464643	464643	472412

Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the supplier-year level. The number of reported observations is the number of non-singleton observations.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

digit sector as the supplier that the firm is being cut off from (at the time of the integration). This means that the coefficient on the "integration with competitor" variable now captures the average sales response for a firm that does not already have any "alternative" suppliers already in place in the sector where its supplier vertically integrates.

Columns (3) and (4) of Table 3.13 show that the point estimates of this coefficient are large and negative: firms that are cut off from a supplier that they do not already have an existing alternative to are suffering a large drop in sales. On the other hand, the presence of alternative suppliers mitigates the sales impact. Note that the sales loss may capture both a movement along the demand curve due to higher variable costs, as well as a potential loss of market share due to the competitor experiencing cost reductions after the vertical integration. At the same time, we see the sales drop only when a supplier vertically integrates with a competitor – so unless the cost reductions are particularly taking place in vertical integration episodes with the buyer's competitors, it is unlikely that this channel plays a major role in driving the buyer's sales response. Columns (5) and (6) show IV estimates where the vertical integration dummy is instrumented by a dummy that is one iff the vertical merger happens up to and including two periods after a mutual fund outflow event. Estimates are very similar to the OLS estimates. In

all specifications the model fit is very good – but that is due to the fixed effects absorbing most of the variation in sales. In Appendix 3.B.1 we show results with employment on the left-hand side. We do not find a drop in firm employment when a supplier integrates with a competitor.

Figure 3.7 shows an event study graph around the time of vertical integration of a supplier with a non-competitor (dashed red line) and with a competitor, for firms that have no existing alternative suppliers (solid blue line). We see that in cases where the supplier is vertically integrating with a competitor, firms' sales are substantially lower if they do not have existing alternative suppliers.

Table 3.13: Impact on buyer's sales

	Dependent variable: Log sales					
	(1)	(2)	(3)	(4)	(5)	(6)
Supplier v.integrates	0.042**	0.018^{+}	0.042**	0.018^{+}	0.020^{+}	0.020^{+}
	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Supplier v. integrates w. competitor	-0.038	-0.052^{+}	-0.137*	-0.143*	-0.036	-0.114*
	(0.031)	(0.030)	(0.060)	(0.058)	(0.029)	(0.046)
$\times$ $log(1 + \#$ alt. suppliers)			0.043*	0.040*		0.035*
			(0.017)	(0.017)		(0.015)
Buyer FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry $\times$ Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls		Yes		Yes	Yes	Yes
Method	OLS	OLS	OLS	OLS	IV	IV
Observations	77202	77202	77202	77202	77202	77202
R^2	0.98	0.98	0.98	0.98	0.98	0.98

Note: Controls: number of customers, competitors and suppliers. In columns (5) and (6), the interaction terms are instrumented by an interaction of the competitor status with a dummy that is one if the vertical integration happens up to two years after a mutual fund outflow event. Robust standard errors, clustered at the firm level, are in parentheses.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

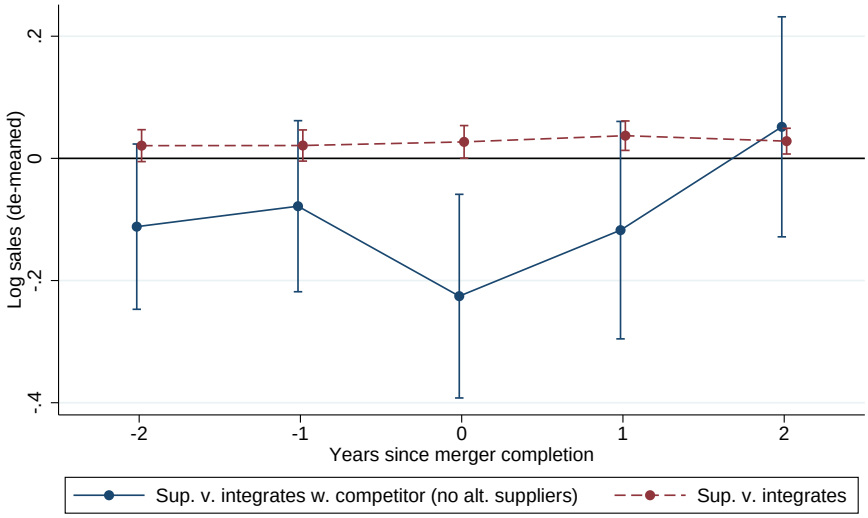
Finally, we look at the sales impact of foreclosure in international mergers. Table 3.14 shows results when we restrict attention to cross-border mergers. The point estimates of the "integration with competitor" dummy are very similar to those in Table 3.13.

3.4.2 Can synergies account for breaking supplier links?

One potential alternative explanation of our finding that vertical relations are more likely to end when the supplier vertically integrates with the buyer's competitor is that there are very strong synergies from the merger. If synergies give the integrated downstream firm a large cost advantage, the unintegrated downstream competitor may be forced to exit the product market, which may lead it to cut its ties to the upstream firm.

If this explanation was driving our results, however, we would expect that vertical integration would adversely affect the market shares of all downstream firms in the industry, including competitors that did not have a supplier relationship with the integrating upstream unit. Table

Figure 3.7: Timing of the correlation of buyers' log sales with vertical integration of a supplier



Note: The figure presents the results of estimating equation 3.4 with two leads and lags for both $\mathbb{1}\{\text{A supplier integrates vertically w. competitor of } b\}_{bt}$ and $\mathbb{1}\{\text{A supplier integrates vertically}\}_{bt}$. Confidence intervals are calculated using robust standard errors clustered at the firm level.

3.15 shows results from a regression of log firm sales on a dummy that is one if the firm has a competitor in that year that vertically integrates (and firm and industry × year fixed effects, as well as the set of controls from above). We find no statistically significant correlation between a competitor vertically integrating and a change in firm sales. This stands in contrast to the situation that we looked at above, where a competitor is vertically integrating with the buyer's supplier, and where we observed a drop in firm sales.

These results are in line with the findings of Blonigen and Pierce (2016), who study the effect of mergers and acquisitions on physical productivity and markups of U.S. manufacturing establishment. They use a similar dataset of public and private mergers and acquisitions, and find no effect of physical productivity of integrating plants, but a significant increase in markups. While their data allows for a much more direct investigation of the productivity effects of mergers and acquisition than our indirect results on competitor's sales, the results support the view that much of the impact of M&A is to reduce competition, and little to increase economic efficiency.

3.4.3 Discussion

Enforcement and welfare

Our results suggest that existing antitrust measures have not managed to fully prevent vertical foreclosure in the sample that we study. Since we find similar results on domestic and international mergers, it does not seem to be the case that international mergers are exposed to more scrutiny from competition authorities. Overall, we get the impression that vertical merger enforcement is lax throughout, possibly because of the intellectual history of the question (Salop

Table 3.14: Sales impact: International Mergers

	Dependent variable: Log sales			
	(1)	(2)	(3)	(4)
Supplier v. integrates (intl. M&A)	0.040**	0.022*	0.041**	0.023*
	(0.012)	(0.011)	(0.012)	(0.011)
Supplier v. integrates w. comp. (intl. M&A)	-0.034	-0.050	-0.097^{+}	-0.108^{+}
	(0.032)	(0.031)	(0.057)	(0.057)
$\times$ $log(1 + \#$ alt. suppliers)			0.027^{+}	0.025
			(0.016)	(0.016)
Buyer FE	Yes	Yes	Yes	Yes
Industry $\times$ Year FE	Yes	Yes	Yes	Yes
Controls		Yes		Yes
Observations	77,202	77,202	77,202	77,202
R^2	0.98	0.98	0.98	0.98

Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the firm level. The number of reported observations is the number of non-singleton observations.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 3.15: Impact of vertical integration on competitors' sales

	Dependent variable: Log sales			
	(1)	(2)	(3)	(4)
A competitor v.integrates	0.013	-0.021		
	(0.040)	(0.039)		
$max(t, t-1)$			0.023	
			(0.035)	
$max(t, t-1, t-2)$				0.056
				(0.036)
Buyer FE	Yes	Yes	Yes	Yes
Industry $\times$ Year FE	Yes	Yes	Yes	Yes
Controls		Yes	Yes	Yes
Observations	120,689	120,689	120,689	120,689
R^2	0.94	0.95	0.95	0.95

Note: The variable in the second (third) row is a dummy that is one if a competitor has undergone a vertical integration in the current or last year (current or last two years). Controls: number of customers, competitors and suppliers. Robust standard errors, clustered at the firm level, are in parentheses. The number of observations is larger here than in Table 3.13 because we have more firms with sales data that have competitor relationships than firms with supplier relationships.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

(2017)).

At the same time, it is likely that there are many other cases of foreclosure than the ones we highlight. Our events are those where a buyer is fully foreclosed, i.e. the relationship fully breaks. A situation that is perhaps more prevalent is where the buyer is facing higher prices offered by the bottleneck supplier. Evaluating such situations would require data on prices.

Even if the vertical foreclosure is taking place in some of the cases we studied, the overall

welfare consequences are not necessarily negative, in particular because consumer prices might fall due to increases in productivity or changes in competition. Frictions in firm-to-firm markets are likely to impose additional transaction costs, which will be reflected in the prices paid by final consumers. A full structural analysis of the welfare cost of vertical foreclosure across a broad range of industries is beyond the scope of this paper, but we view our reduced-form evidence as a first step in this direction.

Mechanism of foreclosure

Having read newspaper coverage and SEC filings related to some of the potentially foreclosing mergers and acquisitions, we find it plausible that in many of these cases the integrating firms are not directly cutting off the competing downstream firm. The documents filed by the integrating firms typically emphasize that existing contracts with customers of the upstream firm will be honored. In some cases, however, firms also state that clauses in these contracts allow the customers to withdraw from the agreement. Upon integration, customers of the integrating supplier may find themselves wanting to break the relationships because continuing the relationship would be associated with a strategic disadvantage on the output market.[19] But even if these customers initiated the break, foreclosure is taking place when they have been hurt by the integration.

As an example, consider the acquisition of hard drive disk platter producer Komag by its customer Western Digital (WD) in 2007. Komag had also been supplying WD's rivals Seagate, Maxtor, and Hitachi, and these relationships ceased after integration. In a conference call with market analysts, a senior executive from WD said about Komag's future relationships with their existing customers: "[...] we are prepared to provide all customers with the committed volumes outlined in their existing volume purchase arrangement. However, customers will determine their [input] requirement. Therefore, there could be a significant reduction in volume from those customers [...]." (Securities and Exchange Commission (2007)).

3.5 Conclusion

This paper presents results that suggest that vertical foreclosure along the extensive margin is occurring among large firms, across a range of sectors in the economy, and both for domestic and international mergers. Vertical relationships are much more likely to break when the supplier is integrating with a competitor of the buyer, than when the supplier is integrating with an unrelated party. Depending on market structure, the firm that integrated with the supplier may have an incentive to prevent its competitor from continuing to purchase from the supplier. We find that this higher hazard rate for links breaking remains statistically significant when only considering integration events that occur after exogenous downward pressure on the

[19] Such as strategic disadvantage may arise through the revelation of information to the competitor. See, e.g. Hughes and Kao (2001) and Chen (2001).

suppliers' stock price. Rumored integration that never takes place is not associated with higher hazard rate. We find that on average firms whose supplier vertically integrated with one of their competitors experience a temporary drop in sales. This sales drop is lower for firms that have relationships with other suppliers from the same industry in place.

References

Alfaro, Laura, Paola Conconi, Harald Fadinger, and Andrew F Newman. 2016. "Do prices determine vertical integration?" *The Review of Economic Studies* 83 (3): 855–888.

Asker, John. 2016. "Diagnosing Foreclosure due to Exclusive Dealing." *Journal of Industrial Economics* 64 (3): 375–410.

Barkai, Simcha. 2016. "Declining labor and capital shares." *Stigler Center for the Study of the Economy and the State New Working Paper Series* 2.

Barrot, Jean-Noel, and Julien Sauvagnat. 2016. "Input Specificity and the Propagation of Idiosyncratic Shocks in Prodction Networks." *The Quarterly Journal of Economics:* 1543–1592.

Bernard, Andrew B, and Swati Dhingra. 2015. *Contracting and the Division of the Gains from Trade.* Technical report. National Bureau of Economic Research.

Bernard, Andrew B, Emmanuel Dhyne, Glenn Magerman, Kalina Manova, and Andreas Moxnes. 2017. "The origins of firm heterogeneity: a production network approach." *Tuck School of Business at Dartmouth, unpublished manuscript.*

Blonigen, Bruce A. 1997. "Firm-specific assets and the link between exchange rates and foreign direct investment." *American Economic Review* 87 (3): 447–466.

Blonigen, Bruce A, and Justin R Pierce. 2016. "Evidence for the Effects of Mergers on Market Power and Efficiency." *NBER Working Paper* 22750.

Boehm, Christoph, Aaron Flaaen, and Nitya Pandalai-Nayar. 2015. "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake."

Bollaert, Helen, and Marieke Delanghe. 2015. "Securities Data Company and Zephyr, data sources for M&A research." *Journal of Corporate Finance* 33:85–100.

Bolton, Patrick, and Michael D. Whinston. 1993. "Incomplete Contracts, Vertical Integration, and Supply Assurance." *The Review of Economic Studies* 60 (1): 121–148.

Breinlich, Holger. 2008. "Trade liberalization and industrial restructuring through mergers and acquisitions." *Journal of international Economics* 76 (2): 254–266.

Bureau Van Dijk. 2017. *Zephyr.*

Carballo, Jerónimo, Gianmarco IP Ottaviano, and Christian Volpe Martincus. 2018. "The buyer margins of firms' exports." *Journal of International Economics* 112:33–49.

Carvalho, Vasco M, Makoto Nirei, Yukiko Saito, and Alireza Tahbaz-Salehi. 2016. "Supply chain disruptions: Evidence from the great east japan earthquake."

Chen, Yongmin. 2001. "On Vertical Mergers and Their Competitive Effects." *The RAND Journal of Economics* 32 (4): 667–685.

Chipty, Tasneem. 2001. "Vertical Integration , Market Foreclosure , and Consumer Welfare in the Cable Television Industry." *American Economic Review* 91 (3): 428–453.

Crawford, Gregory S., Robin S. Lee, Michael D. Whinston, and Ali Yurukoglu. 2018. "The Welfare Effects of Vertical Integration in Multichannel Television Markets." *Econometrica* 86 (3): 891–954.

Cunningham, Colleen, Florian Ederer, and Song Ma. 2018. *Killer acquisitions.* Technical report.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2018. *The rise of market power and the macroeconomic implications.* Technical report.

Dessaint, Olivier, Thierry Foucault, Laurent Frésard, and Adrien Matray. 2016. "Ripple effects of noise on corporate investment."

Edmans, Alex, Itay Goldstein, and Wei Jiang. 2012. "The real effects of financial markets: The impact of prices on takeovers." *The Journal of Finance* 67 (3): 933–971.

Ekholm, Karolina, Rikard Forslid, and James R Markusen. 2007. "Export-platform foreign direct investment." *Journal of the European Economic Association* 5 (4): 776–795.

Giroud, Xavier, and Holger M Mueller. 2017. *Firms' Internal Networks and Local Economic Shocks.* Technical report. National Bureau of Economic Research.

Goldberg, Pinelopi K, Amit K Khandelwal, Nina Pavcnik, and Petia Topalova. 2010. "Multiproduct firms and product turnover in the developing world: Evidence from India." *The Review of Economics and Statistics* 92 (4): 1042–1049.

Guadalupe, Maria, Olga Kuzmina, and Catherine Thomas. 2012. "Innovation and foreign ownership." *American Economic Review* 102 (7): 3594–3627.

Gugler, Klaus, Dennis C Mueller, B Burcin Yurtoglu, and Christine Zulehner. 2003. "The effects of mergers: an international comparison." *International journal of industrial organization* 21 (5): 625–653.

Gugler, Klaus, and Florian Szücs. 2016. "Merger externalities in oligopolistic markets." *International Journal of Industrial Organization* 47:230–254.

Gutierrez, Germán, and Thomas Philippon. 2018. *How EU Markets Became More Competitive Than US Markets: A Study of Institutional Drift.* Working Paper, Working Paper Series 24700. National Bureau of Economic Research, June.

Gutiérrez, Germán, and Thomas Philippon. 2017. *Declining Competition and Investment in the US.* Technical report. National Bureau of Economic Research.

Harford, Jarrad, Robert Schonlau, and Jared Stanfield. 2019. "Trade Relationships , Indirect Economic Links , and Mergers." *Management Science* 65 (July): 3085–3110.

Hart, Oliver, and Jean Tirole. 1990. "Vertical Integration and Market Foreclosure." *Brookings Papers on Economic Activity. Microeconomics* 1990 (1990): 205–286.

Hastings, Justine S., and Richard J. Gilbert. 2005. "Market power, vertical integration and the wholesale price of gasoline." *Journal of Industrial Economics* 53 (4): 469–492.

Hoberg, Gerard, and Gordon Phillips. 2010. "Product market synergies and competition in mergers and acquisitions: A text-based analysis." *The Review of Financial Studies* 23 (10): 3773–3811.

Hortaçsu, Ali, and Chad Syverson. 2007. "Cementing Relationships: Vertical Integration, Foreclosure, Productivity, and Prices." *Journal of Political Economy* 115 (2): 250–301.

Hughes, John S, and Jennifer L Kao. 2001. "Vertical integration and proprietary information transfers." *Journal of Economics & Management Strategy* 10 (2): 277–299.

Huneeus, Federico. 2018. "Production Network Dynamics and the Propagation of Shocks."

Kalemli-Ozcan, Sebnem, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. 2015. "How to Construct Nationally representative Firm Level Data from the Orbis Global Database." *NBER Working Paper,* no. 21558.

Kikkawa, Ayumu Ken, Glenn Magerman, Emmanuel Dhyne, et al. 2018. *Imperfect competition in firm-to-firm trade.* Technical report.

Klein, Benjamin, Robert G. Crawford, and Armen A. Alchian. 1978. "Vertical integration, appropriable rents, and the competitive contracting process." *Journal of Law and Economics* 21 (2): 297–326.

Lafontaine, Francine, and Margaret Slade. 2007. "Vertical integration and firm boundaries: The evidence." *Journal of Economic Literature* 45 (3): 629–685.

Maksimovic, Vojislav, Gordon Phillips, and Liu Yang. 2013. "Private and public merger waves." *The Journal of Finance* 68 (5): 2177–2217.

Malmendier, Ulrike, Enrico Moretti, and Florian S Peters. 2018. "Winning by losing: evidence on the long-run effects of mergers." *The Review of Financial Studies* 31 (8): 3212–3264.

Nocke, Volker, and Stephen Yeaple. 2007. "Cross-border mergers and acquisitions vs. greenfield foreign direct investment: The role of firm heterogeneity." *Journal of International Economics* 72 (2): 336–365.

Oberfield, Ezra. 2018. "A Theory of Input–Output Architecture." *Econometrica* 86 (2): 559–589.

Ordover, Janusz A., Garth Saloner, and Steven C. Salop. 1990. "Equilibrium Vertical Foreclosure." *American Economic Review* 80 (1): 127–142.

Rauh, Joshua D, and Amir Sufi. 2011. "Explaining corporate capital structure: Product markets, leases, and asset similarity." *Review of Finance* 16 (1): 115–155.

Rey, Patrick, and Jean Tirole. 2007. "A Primer on Foreclosure." Chap. 33 in *Handbook of Industrial Organization,* edited by Mark Armstrong and Robert H. Porter, 3:2145–2220. North Holland.

Rhodes-Kropf, Matthew, and Steven Viswanathan. 2004. "Market valuation and merger waves." *The Journal of Finance* 59 (6): 2685–2718.

Salop, Steven C. 2017. "Invigorating vertical merger enforcement." *Yale Law Journal* 127:1962.

Salop, Steven C, and Daniel P Culley. 2015. "Revising the US vertical merger guidelines: policy issues and an interim guide for practitioners." *Journal of Antitrust Enforcement* 4 (1): 1–41.

Securities and Exchange Commission. 2007. *Preliminary Communications on Schedule 14D-9 related to Komag Inc.* https://www.sec.gov/Archives/edgar/data/106040/000089256907000897/ EDGAR File No. 005-39184 07956586.

Shenoy, Jaideep. 2012. "An examination of the efficiency, foreclosure, and collusion rationales for vertical takeovers." *Management Science* 58 (8): 1482–1501.

Slade, Margaret E. 2019. *Vertical Mergers: Ex Post Evidence and Ex Ante Evaluation Methods.* Technical report.

Stiebale, Joel. 2016. "Cross-border M&As and innovative activity of acquiring and target firms." *Journal of International Economics* 99:1–15.

Stiebale, Joel, and Florian Szücs. 2017. *The Effects of Mergers on Markups, Productivity, and Innovation of Rivals.* Technical report. Mimeo, DICE Dusseldorf.

The Economist. 2018. "Regulators across the West are in need of a shake-up," no. 15 November 2018.

Tintelnot, Felix, Ayumu Ken Kikkawa, Magne Mogstad, and Emmanuel Dhyne. 2018. *Trade and domestic production networks.* Technical report. National Bureau of Economic Research.

Waterman, David, and Andrew A. Weiss. 1996. "The effects of vertical integration between cable television systems and pay cable networks." *Journal of Econometrics* 72 (1-2): 357–395.

Appendices

3.A Data sources and definitions

We combine three components to construct the database used in this paper:

- A production and competitor network between large firms from FactSet Revere
- A comprehensive M&A database, Bureau van Dijk's Zephyr, with information on deals and rumors about deals
- Company financials and industry classifications from Bureau van Dijk's Orbis, Compustat and FactSet Fundamentals

This appendix describes each of the data sources as well as the key variables we derive from them.

3.A.1 FactSet Revere supply chain data

Content and data sources

FactSet is a commercial data provider that mainly sells to companies in the financial services sector. Its supply chain data (called "Revere") provides information on the nature and duration of vertical and horizontal relationships between firms. FactSet collects this information on relations from primary public sources such as SEC filings, investor presentations, corporate actions, company websites and press releases. For each firm, FactSet conducts an annual review to update the database. In addition, press releases and corporate actions are monitored daily for US firms.

Each relation between two companies is dated with a start date at which the relation was first recorded by FactSet and with an end date at which it was noticed that the relation no longer existed. In addition, each relation is categorized into buyer links, supplier links, competitor links or partnerships. These broad categories are detailed into 13 subcategories (see Table 3.16). We use these categories to define two types of networks:

- Buyer-supplier network: a directed graph on which an edge is created when the target company is a supplier of the source company, i.e. at least one of the following is true:
  - the source company discloses the target company as a supplier of products or services
  - the target company discloses the source company as a customer of products or services

Table 3.16: Number of relationships in raw FactSet Revere data

	Frequency	Percentage
Supplier	114,136	12.71
Competitor	197,423	21.98
Customer	290,893	32.38
Partner: Distribution	24,725	2.75
Partner: Equity investment	53,602	5.97
Partner: Production	12,737	1.42
Partner: Investor	48,244	5.37
Partner: Joint-Venture	29,845	3.32
Partner: Licensing	37,083	4.13
Partner: Marketing	16,296	1.81
Partner: Other	876	0.10
Partner: Research Collaboration	46,273	5.15
Partner: Technology	26,189	2.92
Total	898,322	

Note: Frequency table of the raw number of relations in the relationship dataset from which we construct the firm network. In line with the description in the documentation of the data, we count companies providing paid distribution, production, marketing and licenses as suppliers.

  - the target company provides paid manufacturing, distribution or marketing services to the source company
  - the target company licenses products, patents, technology or IP to the source company

- Competitor network: an undirected graph on which an edge is created if at least one of the two company discloses the other one as a competitor

We do not include the partnership links provided by fact set for our analyses (Joint ventures, Equity stakes, research collaborations and integrated product offerings).

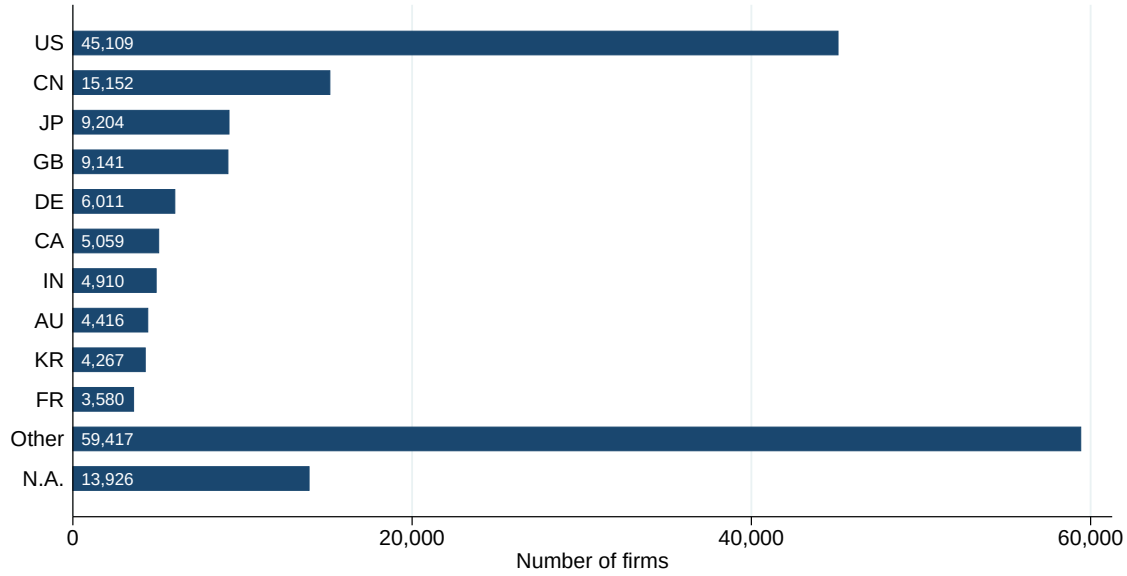
Finally, many relations are also provided with a few keywords explaining the links, though not in a fully systematic fashion. Companies can have multiple links, for instance in order to document that a supplier is also in competition with a given customer.

Coverage

The data contain observations on 180,192 firms, some of which are "covered" companies (in the sense that the data provider actively searches for information on these firms); the others show up as suppliers, buyers, or competitors of covered companies. FactSet determines coverage mainly based on membership of firms in major stock indexes. The provider aims to cover all companies listed in a set of global indexes, such as the FTSE Global All Cap, Russel Global, Stoxx Global and a range of global MSCI indexes. In addition, all US-based publicly traded firms are covered, as well as companies that are part of multiple local and regional stock market indexes, i.e. large non-US multinationals. FactSet achieves high but not complete coverge of the indexes. For example, 90.3% of the firms in MSCI ACWI All Cap have relationship information, 95.4% of

the S&P 500 and 94.5% of the Russell 3000. While these coverage rules favor large listed firms, there are many smaller and non-listed firms in our sample because they have relationships with large firms.

Figure 3.8: Number of firms by country



Note: The figure reports the number of firms in the FactSet database by country of headquarter.

Coverage varies by country. Figure 3.8 reports the number of firms in the database by the country of their headquarters. Consistent with the fact that FactSet originally only covered US firms, about a quarter of the firms is based in the US. Due to efforts to expand the database internationally starting in 2007, and because of foreign firms trading with US firms, international coverage goes well beyond large multinationals.

While the database is not representative even of the universe of US firms, it does contain a wide range of industries. Figure 3.9 reports the number of firms in the sample by a high-level aggregation of NAICS industry codes. The manufacturing sector contains the largest share of firms in our data, followed by financial services and insurances, and then by professional services.

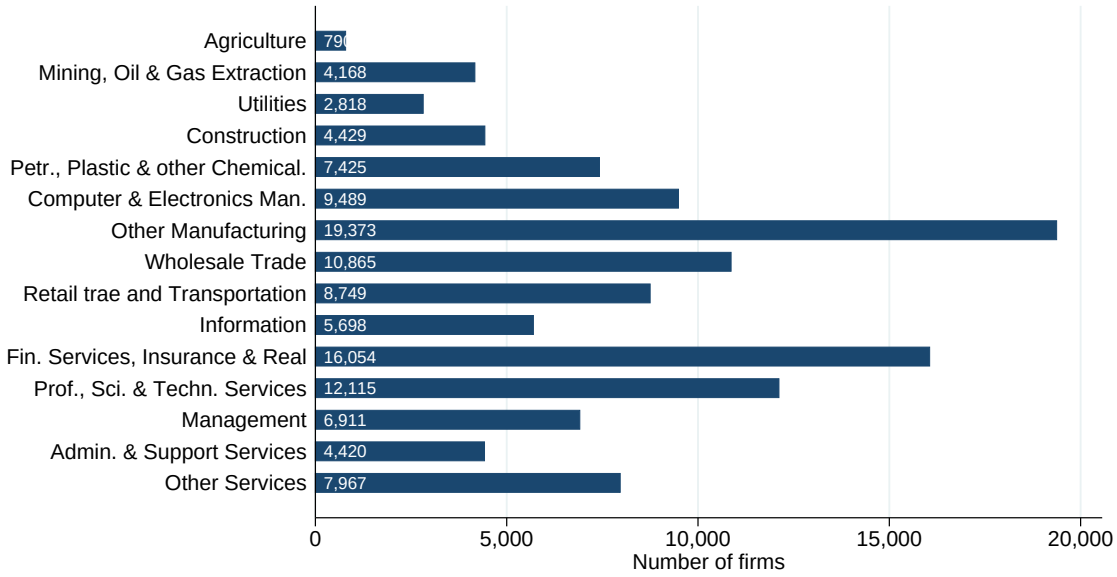
When a company entity in the data ceases to exist, FactSet documents the reasons for it, along with a successor company where it exists. This fact allows us, in particular, to identify the successor company in the case of a complete merger or acquisition so that links are not mechanically breaking at acquisition.

The data start in 2003 and have been gradually expanded over time (Figure 3.10). Non-US firms were included from 2007 onwards.

Key variables

We annualize the relationship data in order to facilitate the matching with the company financials. A relation of any kind is counted as active in a given year if there is at least one day

Figure 3.9: Number of firms by industry



Note: The figure reports the number of firms in the FactSet database by primary industry classification.

between start date and end date of the relation that falls into that year. The result is a panel of relations that is identified by source company, target company and year.

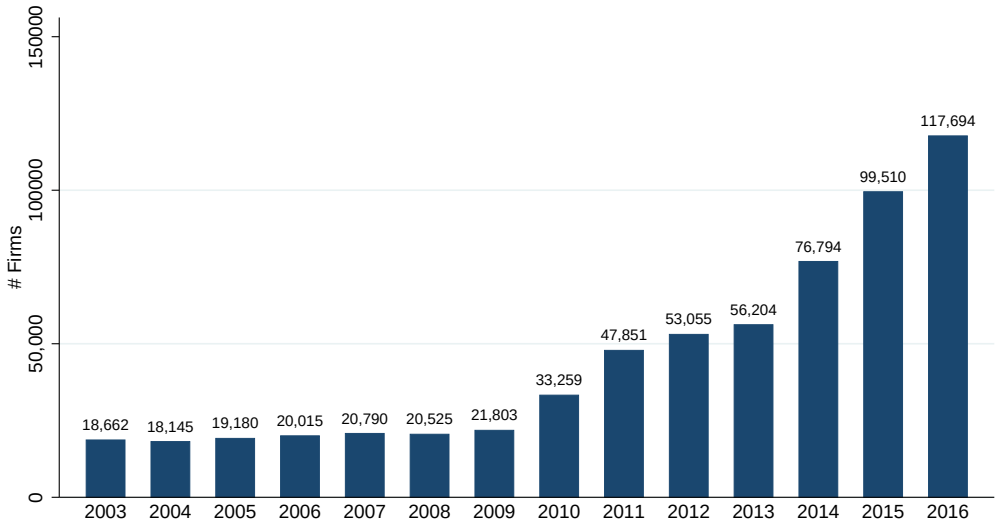
Buyer-supplier link breaks: The variable is one if and only if (i) the relation was active in the previous year but is no longer active and (ii), in case buyer and/or supplier were involved in a merger or acquisition, there is no active link between the successor company or acquiror and the buyer or supplier. The second condition rules out purely mechanically breaks in the supply chain that could result from mergers and acquisitions. This variable is the main left-hand side variable in the regressions in Section 3.3.

We set this variable to missing in a few cases to avoid other possible mechanical breaks. If a buyer has dropped out of coverage and, in case of a merger or acquisition, the successor company or acquiror is not covered by FactSet in the current year, then its relations are not counted as breaking. This is to rule out that we erroneously count a link as broken purely because a firm is no longer covered. We also count the variable as missing when the buyer and the supplier in the given relation are integrating.

3.A.2 Zephyr M&A data

We use Bureau van Dijk's Zephyr database for information on mergers and acquisitions. Zephyr records deals and rumors about deals for mergers and acquisitions in which at least a 2% stake in the target company changes owners and the value of the deal exceeds GBP 1M. For an overview of Zehphyr's content, coverage, and how it compares to other M&A databases, see Bollaert and Delanghe (2015). For the sake of brevity, we refer to any merger or acquisition simply as merger in the following.

Figure 3.10: Number of firms over time



Note: The figure presents the number of firms in the FactSet database by year.

Matching and merging with other data sources

Zephyr reports the exact dates of rumors, announcements and (expected) completions or withdrawals of mergers. Analogously to the FactSet data, we convert these data to a panel of merger events, where each observation is identified by the target firm, the acquiring firm and the calendar year of the completion date for completed mergers or the year of the rumor for mergers that were rumored but never completed.

We match firms in the FactSet and Zephyr databases using security identifiers such as CUSIP or ISIN, as well as ticker names wherever possible. For the remaining firms we use a string matching tool provided by Bureau van Dijk that takes into account company names and, where available, addresses.

Table 3.17: Types of mergers and acquisitions

	Vertical		Horizontal		Both		Unrelated		Total	
	Count	%	Count	%	Count	%	Count	%	Count	%
Partial acquisitions	46	5.8	51	6.4	8	1.0	694	86.9	799	100.0
Full acquisitions & mergers	132	4.4	568	18.9	69	2.3	2,231	74.4	3,000	100.0
Total	178	4.7	619	16.3	77	2.0	2,925	77.0	3,799	100.0

Note: Number of partial and full mergers and acquisitions by presence vertical and horizontal relation between the merging parties (2003-2016). Partial acquisitions exclude minority stakes.

Table 3.17 breaks down the mergers and acquisitions between firms in the matched sample by the type of their relation in the FactSet Revere data. In addition to vertical mergers, the data allow us to identify horizontal mergers (through competitor relationships) and mergers that are both horizontal and vertical in nature. In our analyses, however, we focus on integrations that have a vertical dimension to them.

Table 3.18 reports summary statistics about buyer-supplier relations where the supplier was

Table 3.18: Treated buyer-supplier relations and placebo counterparts

	Vertical M&A with Comp.		
	Actual M&A	Rumored M&A	Difference
New relationships	0.10	0.14	0.04
Ending relationships	0.51	0.16	-0.35***
Buyer's suppliers	71.00	78.09	7.10
Supplier's buyers	27.55	48.19	20.64***
Buyer's competitors	68.24	68.28	0.04
Supplier's competitors	19.58	36.14	16.56***
Age of relationship	3.51	3.98	0.47
Sales (log m$): Buyer	8.26	8.96	0.70**
Sales (log m$): Supplier	6.92	8.77	1.86**
Sales (log m$): Competitor	9.33	9.78	0.45**
Log Employment: Buyer	9.23	9.84	0.61*
Log Employment: Supplier	6.79	9.46	2.68**
Log Employment: Competitor	10.32	9.86	-0.46*
N	207	221	428

Note: Summary statistics for the buyer-supplier-years for which suppliers are involved in a vertical M&A-transaction with a competitor of the buyer.

vertically integrating or rumored to be vertically integrating with a competitor of the buyer. While the buyers in both groups are quite comparable, it seems that rumors involve suppliers that are on average somewhat larger than those suppliers which actually undergo integration. Note that we control for these differences in our regressions and also do not find a differential effect for larger or smaller suppliers.

Key variables

With the firm network and the merger information in place, we construct our main explanatory variables. For ease of exposition, for a given buyer-supplier-year observation, we refer to the buyer as firm b, to the supplier as firm s and to a firm that merges with the supplier as firm c.

Supplier vertically integrates: We construct a dummy variable that is equal to one at the buyer-supplier-year level if firm s is involved in a merger with firm c which is also a customer of s. We restrict attention here to full mergers and acquisitions in the sense that the stake of the acquirer after the acquisition is 100% but was either zero or unknown before. Firm s can be either the acquirer or the target in the M&A with firm c. Note that we only count mergers as vertical if there was an active buyer-supplier relationship between s and c in the year of integration.

Supplier vertically integrates with buyer's competitor: This dummy variable is equal to one at the buyer-supplier-year level if firm s and c are merging, s is an active supplier of c in that year and b and c have an active competitor relationship in that year.

For the placebo analyses we construct the same variables again using rumored mergers instead of actual mergers. These rumors come from "unconfirmed reports", which "may be in the press, in a company press release, or elsewhere" (Bureau Van Dijk (2017)). They may

indeed come from announcements by one of the involved firm as long as the other firms have not yet confirmed the announcements. In the Zephyr database, this corresponds to deals for which the variable *deal status* is "Rumour". The timing of these events differs slightly: instead of the completion date (which is unavailable), we use the rumor date. In general however, there is little time elapsing between a rumor and the completion of a deal: 145 days on average and about 92% of rumors which turn out to be true are realized within a year. For our placebo analyses, we exclude all rumors that materialize within three years.

3.A.3 Company financials and industry classifications

To achieve best possible coverage of company financials and industry classifications for the firms in our supply chain network, we combine data from Orbis, Compustat (through WRDS) and FactSet Fundamentals. The combination of the various data sources is necessary in particular because of varying coverage over time. While we have supply chain and merger information available from 2003 to present, Orbis data is only available to us from 2007 onwards. In contrast, Compustat and FactSet Fundamentals are available for earlier years as well.

Matching and merging with other data sources

As with the Zephyr database, we first match all firms for which securities identifiers are available. As Zephyr and Orbis share the same identifier, matching these data sources is straightforward. For the remaining firms and data sources we use the company names for string matching.

For firms where financials are available from multiple data sources, we only retain the information from the data source that provides the longest coverage of the sales variable of that firm. Hence, all of a given firm's financial information always come from the same data source in order to ensure consistency over time and across items. Wherever ties occur, preference is given first to FactSet Fundamentals, then to Orbis. Note that the variables from several datasets are almost perfectly correlated for the observations where we do have overlaps in coverage.

Key variables

Sales: The sales data are contained in the variables "ff_sales" in FactSet Fundamentals, "sales" in Orbis and "sale" in Compustat. Orbis reports all financials directly in USD, the sales data from the other data sources are converted to USD where necessary using exchange rate information included in those datasets. A few firms in the data exhibit unusual sales trajectories that seem to suggest reporting or data entry issues. In order to rule out that our results are driven by such observations, we exclude firms whose sales growth falls into the first or 99th percentile in one or more years.

Employment: The number of employees is contained in the variables "ff_emp" in FactSet Fundamentals, and "emp" in Orbis and Compustat. We use these variables without further processing.

NAICS codes: From Orbis and Compustat we can also retrieve NAICS industry codes ("naics_primary" and "naics_secondary" in Orbis, "naics" in Compustat). When several NAICS codes are available, we restrict attention to the primary one for clustering or aggregation.

Mutual fund capital outflow instrument

To construct the MFHS instrument, we follow Appendix C of Dessaint et al. (2016). We construct quarterly capital net outflows of US mutual funds using the CRSP mutual funds data, and the hypothethical stock sales following large outflows using the funds' portfolio data in CDA Spectrum/Thomson. We match funds using the crosswalk provided by WRDS.

3.A.4 Summary statistics for vertically integrating firms

Table 3.19 shows summary statistics for firms that vertically integrate with one of their buyers. The left column contains firms that integrate with non-competitors of one of their buyers, the right column contains firms that integrate with a firm that is not in a competitor relationship with any other buyer.

Table 3.19: Summary statistics for vertically integrating firms

	No foreclosure potential		Foreclosure potential	
	Mean	SD	Mean	SD
# Buyers	47.81	70.72	18.51	15.29
# Suppliers	46.54	66.66	15.38	16.81
# Competitors	48.59	67.07	17.57	12.50
Log sales	15.50	2.04	–	–
Log employment	9.37	2.23	–	–
Observations	140		53	

Note: The table presents summary statistics on suppliers that vertically integrate with a non-competitor of the buyer (first column) or a competitor of the buyer (second column). Coverage of sales and employment is very poor (<10%) for suppliers in vertical integrations with foreclosure potential.

3.B Further results

3.B.1 Impact of foreclosure on employment

Table 3.20 shows the impact of a supplier integrating with a competitor on firm employment. The OLS results are similar to sales, though somewhat smaller (about half of the percentage-wise effect on sales) and not statistically significant. IV estimates are not significantly different from zero either. Overall, there does not seem to be an impact of foreclosure on employment.

Table 3.20: Impact on buyer's employment

	Dependent variable: Log employment				
	(1)	(2)	(3)	(4)	(5)
Supplier v.integrates	0.028**	0.009	0.029**	0.009	0.008
	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
Supplier v. integrates w. competitor	-0.014	-0.028	-0.077	-0.086	0.035
	(0.053)	(0.052)	(0.100)	(0.098)	(0.184)
$\times$ $log(1 + \#$ alt. suppliers)			0.027	0.025	0.012
			(0.024)	(0.024)	(0.035)
Buyer FE	Yes	Yes	Yes	Yes	Yes
Industry $\times$ Year FE	Yes	Yes	Yes	Yes	Yes
Controls		Yes		Yes	Yes
Method	OLS	OLS	OLS	OLS	IV
Observations	70983	70983	70983	70983	70983
R^2	0.98	0.98	0.98	0.98	0.98

Note: Controls: number of customers, competitors and suppliers. Robust standard errors, clustered at the firm level, are in parentheses.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

3.B.2 Direct comparison of the rumored vs actual mergers

Table 3.21 shows regressions with both rumored and actual integration events included. This allows for a direct comparison of the two types of events in the same specification. While the main coefficient of interest – the impact of an actual vertical merger of the supplier with the buyer's competitor – remains large and significant, the rumored counterparts of these events have slightly negative and statistically not significant coefficients.

3.B.3 Hazard models

An alternative way to model the impact of a supplier's vertical integration on the probability of a buyer-supplier link breaking is in terms of a hazard model. In this framework, we can compare the survival times of links where the supplier integrated with a competitor of the buyer to the survival times of links where there was a vertical integration with a non-competitor. When a link does not break during the observation period, the survival time is treated as censored.

Table 3.22 presents the result of a Cox proportional hazard model estimated on all buyer-supplier links in which the supplier vertically integrated with one of its customers:

$$h_{bs}(t) = h_{0bs}(t) \exp\left(\beta \mathbb{1}\{s \text{ integrates vertically w. competitor of } b\}_{bs} + \eta_b + \eta_t + \eta_{i(b)} + \eta_{i(s)}\right)$$

where $h_{bs}(t)$ is the hazard of a buyer-supplier link breaking and η_b , η_t, $\eta_{i(b)}$ and $\eta_{i(s)}$ are indicator variables for buyers, integration years, the buyer's industry and the supplier's industry respectively. We use a partial likelihood framework that does not require to specify the baseline

Table 3.21: Comparison of rumored and actual vertical mergers with competitor of the buyer

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$		
	(1)	(2)	(3)
Supplier v. integrates	0.014	0.006	0.021
	(0.020)	(0.020)	(0.019)
Sup. v. integrates, rumor about competitor	-0.027	-0.020	-0.038
	(0.043)	(0.074)	(0.067)
Supplier v. integrates w. competitor	0.208**	0.199*	0.186*
	(0.071)	(0.094)	(0.082)
Controls			Yes
Relation FE	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes
Industry Pair × Year FE		Yes	Yes
R^2	0.578	0.619	0.671
Observations	640708	472763	472763

Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the supplier-year level. The number of reported observations is the number of non-singleton observations.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

hazard $h_{0bs}(t)$.[20] The survival time is measured from the beginning of the year of integration to the end of the year in which the relationship broke. Columns (1) to (3) retain all integration events, even when a given buyer-supplier link undergoes multiple vertical mergers of the supplier. Columns (4) to (6) only keep the first of these events for each buyer-supplier link. We calculate standard errors clustered among observations related to the same vertical integration event.

The results confirm those presented in Section 3.3. Even after adding controls and including buyer, year and industry dummies, a vertical relationship is expected to end 28% sooner when the supplier vertically integrates with one of the buyer's competitors compared to when vertical integration occurs with a non-competitor of the buyer. The mean survival time of relations is 4.3 years. While the point estimates vary slightly, the qualitative result that links threatened by vertical foreclosure are substantially more short-lived is robust to choosing alternative specifications of the model and the underlying sample.

3.B.4 Alternative relationship definitions

As an additional robustness check, we repeat the baseline regressions with an alternative definition of the buyer-supplier network. In particular one might be worried that relationships break only temporarily and are ultimately reestablished so that the effects we attribute to vertical foreclosure are only transitory. To address this concern, we can remove gaps from buyer buyer-supplier relations. We count a relation as active in a given year if it has been reported active

[20] The results are similar when specifying an exponential or Weibull distribution.

Table 3.22: Impact of supplier's integration with buyer's competitor on hazard rate of buyer-supplier link breaking

	Hazard ratio of buyer-supplier link breaking					
	(1)	(2)	(3)	(4)	(5)	(6)
Supplier v. integrates w. competitor	1.445**	1.250*	1.283**	1.412**	1.216*	1.228*
	(0.120)	(0.116)	(0.123)	(0.132)	(0.115)	(0.121)
Controls		Yes	Yes		Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Buyer Industry dummies			Yes			Yes
Supplier Industry dummies			Yes			Yes
Events	All	All	All	First	First	First
R^2						
Observations	6934	6934	6223	5456	5456	4871

Note: Controls: number of upstream customers and competitors, age of the link. Robust standard errors clustered at the supplier-year level.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

in previous and in future years, even when it is currently not reported to be active. Here, a link is not counted as breaking if it reforms at a later date.

Table 3.23: Extensive margin regressions with stable buyer-supplier relations

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$			
	(1)	(2)	(3)	(4)
Supplier v. integrates	0.012	0.008	0.006	0.011
	(0.008)	(0.007)	(0.007)	(0.007)
Supplier v. integrates w. competitor		0.186**	0.204**	0.186**
		(0.046)	(0.050)	(0.045)
Controls				Yes
Relation FE	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes
Industry Pair × Year FE			Yes	Yes
R^2	0.434	0.434	0.451	0.492
Observations	3351207	3351188	2521097	2521097

Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the supplier-year level. The number of reported observations is the number of non-singleton observations. The drop in the number of observations in columns (3) and (4) is explained by firms with missing industry codes.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Figure 3.23 shows that the baseline results are robust to using this alternative definition. The coefficients on the variable representing a merger of a supplier with the buyer's competitor remains statistically significant throughout and is even slightly larger than in the baseline results. This means that the correlations are not driven by brief pauses in a relationship which ultimately resumes.

3.B.5 Firm entry in the upstream segment?

Do vertical mergers with foreclosure potential deter entry of firms in the upstream segment? To investigate this question we count the number of suppliers that enter an industry as measured by three-digit NAICS cells in a given year and relate it to whether or not there has been a potential foreclosure event in that industry. This exercise is more tentative in nature for three main reasons: First, NAICS codes are a relatively crude measure of the upstream market that does not account for product or geographical differentiation. Second, the industry codes reported in our data are time invariant, meaning that we cannot capture existing firms moving into new product markets. Finally, as we document above, our sample consists mainly of large firms and we may therefore not be able to detect changes in the entry patterns of small firms.

Notwithstanding these caveats, we estimate the following regression:

$$LogEntry_{it} = \beta \mathbb{1}\{Potential\ foreclosure\}_{it} + \eta_i + \eta_t + \varepsilon_{it}$$

where $LogEntry_{it}$ is defined as the log of one plus the number of suppliers in industry i that have at least one customer in year t but did not have one the previous year.

Table 3.24 reports the results of this regression. We measure potential foreclosure in two different ways. The first approach is a dummy indicating whether a supplier in the industry had a merger with foreclosure potential, i.e. the supplier merged with a buyer whose competitor it also supplied (columns 1-2). The second approach is a dummy indicating that a merger with foreclosure potential coincided with a break of the buyer-supplier link with the downstream competitor (columns 3-4). While the estimates are quite noisy and not statistically significant, the point-estimates are negative throughout. Perhaps foreclosure events have a negative impact on firm entry in the upstream industry. Because of the caveats mentioned above and the fact that estimates are not very precise, we hesitate to draw conclusions.

Table 3.24: Firm entry and vertical integrations with foreclosure potential

	Dep. var.: $log(1 + \#entering\ suppliers)_{it}$			
	(1)	(2)	(3)	(4)
V. integration w. foreclosure potential	-0.032		-0.079	
	(0.111)		(0.092)	
× buyer-supplier link breaks		-0.045		-0.110
		(0.130)		(0.107)
Controls			Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
R^2	0.891	0.891	0.926	0.926
Observations	1236	1236	1236	1236

Note: Controls: number of buyer-supplier relations and number of suppliers in a given industry-year.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

3.B.6 Relationship severance prior to integration?

We revisit the main regression presented in section 3.3 in order to investigate whether buyer-supplier relations are more likely to end already in the year before the supplier vertically integrates. If there was reverse causality for supply assurance reasons for instance, we might expect to find a positive correlation in the year before integration actually takes place. We repeat the baseline specification but replace the right-hand side variables by dummies that are one if a time-t competitor of b vertically integrates with s at time $t+1$.

Table 3.25 presents the results of this exercise. We find no evidence that relationships are already more likely to break in the year preceding vertical integration. Quite to the contrary, these relationships are substantially less likely to break. This correlation is not mechanical and in particular persists when restricting the sample to relations with suppliers that keep at least one customer in year $t+1$ when the integration takes place.

Table 3.25: Hazard of links breaking in year before vertical integration

	Dependent variable: $\mathbb{1}\{\text{LinkBreaks}\}_{bst}$			
	(1)	(2)	(3)	(4)
Supplier v. integrates in t+1	-0.031	-0.022	-0.008	0.008
	(0.020)	(0.020)	(0.019)	(0.019)
Supplier v. integrates w. competitor in t+1		-0.186**	-0.174**	-0.151**
		(0.041)	(0.043)	(0.043)
Controls				Yes
Relation FE	Yes	Yes	Yes	Yes
Buyer × Year FE	Yes	Yes	Yes	Yes
Industry Pair × Year FE			Yes	Yes
R^2	0.576	0.576	0.616	0.670
Observations	638681	638681	470788	470788

Note: Controls: number of upstream customers and competitors, age of the link, dummy indicating other links of the supplier breaking. Robust standard errors clustered at the supplier-year level. The number of reported observations is the number of non-singleton observations.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

List of Figures

List of Tables

Institut d'Etudes Politiques de Paris
Ecole Doctorale de Sciences Po
Programme doctoral en économie
Département d'Economie
Doctorat en sciences économiques

Trois Essais en Microéconomie Appliquée: Normes et Réseaux

Résumé

Jan Sonntag

Directeur de thèse:
Emeric Henry, Associate Professor au Département d'Economie de Sciences Po
Soutenu le 12 décembre 2019

Membres du jury:

Charles Angelucci	Assistant Professor, Columbia Business School
Johannes Boehm	Assistant Professor, Sciences Po
Béatrice Boulu-Reshef	Professeur des Universités, Université d'Oréeans *(rapportrice)*
Maria Guadalupe	Professor, INSEAD *(rapportrice)*
Emeric Henry	Associate Professor, Sciences Po *(directeur de thèse)*

Résumé en français

Les trois parties de cette thèse s'articulent autour de deux thèmes : les normes sociales et les réseaux de production. Ces deux thèmes se caractérisent par la présence d'importantes externalités, ce qui en fait des domaines de recherche économique à la fois fascinants et pertinents pour la conception de mesures incitatives et de politiques publiques. Comme les deux sujets sont par ailleurs très différents, je vais les décrire l'un après l'autre.

Les deux premiers chapitres ont pour but d'approfondir notre compréhension de la manière dont les normes sociales façonnent notre comportement. Il existe aujourd'hui une abondante littérature qui montre l'importance des normes sociales pour expliquer le comportement humain dans un large éventail de situations, tel que la participation au marché du travail, le travail d'équipe et la fourniture d'efforts, la participation politique, les dons et autres formes de comportements prosociaux, l'effort appliqué à l'école et au travail, pour n'en citer que quelques-uns. Les normes sociales sont déjà utilisées dans de nombreux contextes pour amener les gens à adopter un comportement souhaité et surmonter les problèmes d'action collective, par exemple pour les inciter à aller voter ou à économiser l'électricité (Allcott (2011) and Gerber and Rogers (2009)). En effet, les interventions utilisant des normes ont souvent l'atout d'être moins coûteuses et plus faciles à mettre en œuvre que les incitations financières.

Un manque de compréhension de la manière dont les normes se forment, dont elles sont soutenues et dont elles façonnent le comportement des individus peut conduire à des inefficacités coûteuses au travers de politiques ou d'institutions mal conçues. Loin d'être une réflexion après coup, les normes sociales peuvent même inverser l'efficacité des incitations monétaires et ainsi faire retourner des mesures d'incitation bien intentionnées contre elles-mêmes. En outre, il a été démontré que les individus sont prêts à punir autrui pour maintenir ces normes, mêmes quand ces punitions sont coûteuses pour eux-mêmes. La recherche sur les normes sociales permet donc d'anticiper ces types de comportements et d'en tenir compte dans la conception des contrats, des institutions et des politiques.

Dans les modèles économiques contemporains, les normes sociales sont généralement conceptualisées en maintenant l'hypothèse d'agents rationnels et en modifiant leurs fonctions d'utilité. Par exemple, les agents en Bénabou and Tirole (2006) dérivent de l'utilité d'un comportement conforme aux normes sociales par deux voies : Premièrement, il y a la motivation intrinsèque d'un agent à se comporter d'une certaine manière, une composante directe de la fonction d'utilité qui entraîne un gain associé à un comportement. Deuxièmement, il y a l'image sociale, le goût pour le fait d'être jugé de manière favorable. La préoccupation de leur image peut en effet inciter les individus à se comporter de manière beaucoup plus prosociale que si leur comportement ne leur donnait pas le bénéfice de l'admiration des observateurs ou le poids de leur désapprobation.

Ce cadre de réflexion sur les normes sociales motive mes deux premiers chapitres. Le premier

porte sur une étude de cas contemporaine où les normes sociales sont utilisées dans la lutte contre le discours haineux en ligne. Je montre que le fait de dénoncer les opinions haineuses est un moyen efficace de dissuader d'autres discours haineux. Le mécanisme qui explique le mieux cet effet est que cette forme de contradiction véhémente sert de punition non-monétaire qui communique la présence d'une norme sociale ou en accentue l'importance. Au-delà de son application directe à la lutte contre les discriminations, ce chapitre peut nous éclairer sur la façon dont les normes influencent le comportement politique plus généralement.

Le deuxième chapitre porte sur le rôle essentiel que joue le goût pour l'image sociale pour expliquer l'effet des normes sociales sur le comportement. En effet, alors que de nombreuses études montrent que ces goûts affectent le comportement des gens en moyenne, nous ne savons pas encore quels individus sont les plus susceptibles d'adapter leur comportement en fonction de l'image sociale. Il est cependant important de prendre l'hétérogénéité du goût pour l'image en compte pour comprendre les effets de l'image sociale et ses conséquences distributives. Notre manque de connaissance dans ce domaine s'explique en grande partie par le fait qu'il n'existe pas de méthode standard pour mesurer le goût pour l'image social au niveau individuel. Le deuxième chapitre présente une expérience novatrice conçue pour combler ce vide. Il permet de calculer une mesure individuelle de préoccupation pour l'image, montre qu'il y a une hétérogénéité substantielle même dans un petit échantillon et analyse sa corrélation avec d'autres préférences sociales.

Le dernier chapitre de ma thèse étudie les réseaux de production. Dans l'économie moderne, la division du travail s'est étendue des travailleurs aux entreprises, de sorte que les chaînes de production forment des réseaux complexes qui s'étendent sur toute la planète. Il est important de comprendre les caractéristiques et les déterminants de ces réseaux, à la fois parce que leur structure influence à quel point les chocs locaux se propagent en fluctuations agrégés et parce qu'ils affectent la performance des entreprises.

L'un des paramètres clés de ces réseaux est le périmètre des entreprises : les chaînes de production sont caractérisées par des degrés variables d'intégration verticale. Si la spécialisation peut offrir des gains d'efficacité, il en va de même de l'exploitation des synergies entre entreprises verticalement liées, de l'élimination de la double marginalisation ou de l'atténuation du problème du « hold-up ». Toutefois, l'intégration verticale peut également donner lieu à des comportements anticoncurrentiels. Dans le dernier chapitre, j'aborde l'un de ces comportements, appelé verrouillage vertical, par lequel les entreprises verticalement intégrées coupent l'approvisionnement de leurs concurrents en intrants essentiels. J'utilise de nouvelles données sur les réseaux de production pour identifier les fusions et acquisitions entre entreprises verticalement liées et évaluer dans quelle mesure ces fusions affectent les chaînes d'approvisionnement de leurs concurrents.

Malgré leur apparente disparité, les trois chapitres partagent quelques caractéristiques importantes. Tous les trois sont motivés par une littérature théorique existante mais sont principalement de nature empirique. Les trois chapitres reposent sur des efforts distincts de collecte de données dans le but d'apporter un nouvel éclairage à des problèmes de longue date. En

effet, chaque chapitre contient des éléments qui pourront servir de base à d'autres recherches : le premier chapitre montre comment on peut avoir recours dans notre discipline à des techniques de « deep learning » encore peu utilisées en économie, le deuxième chapitre décrit une expérience pour mesurer des problèmes d'image qui peuvent facilement être inclus dans d'autres expériences et les données sur les réseaux de production au cœur du troisième chapitre pourraient également être utiles pour examiner des questions autres que la verrouillage vertical.

Je procède par résumer chacun de ces chapitres plus en détail.

1er chapitre – Normes sociales et réseaux sociaux

Suite à une vague de victoires électorales de populistes de droite dans de nombreux pays occidentaux, la capacité des réseaux sociaux à influencer les opinions et les actes des individus a été vivement débattue au cours des dernières années. Un sujet d'inquiétude particulier dans ce contexte est la propagation de discours haineux. Ceci est d'autant plus inquiétant qu'il y a maintenant de plus en plus de preuves empiriques établissant un lien causal entre discours haineux et crimes contre ceux qui sont ciblés par ces discours : les mots inspirent les actes.

Dans le premier chapitre de ma dissertation, je m'interroge sur la question des interventions basées sur les normes sociales : sont-elles capables de limiter la prévalence de messages haineux et donc leurs conséquences néfastes? Une approche qui pourrait établir ou communiquer des normes définissant le type de propos considéré comme acceptable est le « contre-discours », une intervention qui consiste à contredire des messages haineux. Ce type d'intervention décentralisée est-elle efficace? Plus généralement, dans quelle mesure s'influencent mutuellement les individus dans les débats en ligne?

Afin de répondre à ces questions, j'utilise les pages Facebook de grands médias en allemand comme laboratoire. Je considère un groupe de 35 000 utilisateurs qui organise des interventions de contre-discours. Ce groupe se coordonne pour intervenir sur 1 ou 2 articles de presse sur Facebook par jour qui ont été particulièrement ciblés par les messages haineux. Lors de ces interventions, les membres écrivent des commentaires pour condamner les messages haineux sur l'article en question et contredisent directement les messages haineux postés par d'autres utilisateurs. Ceci est très visible pour d'autres utilisateurs : du fait de sa taille importante et du nombre limité d'interventions, le contre-discours du groupe représente souvent une partie importante des commentaires sur les articles ciblés.

Durant six mois, j'ai enregistré tous les articles publiés par des grands médias allemands sur Facebook, ainsi que toutes les réactions publiques des utilisateurs. J'infère quels messages contiennent des discours haineux en utilisant un réseau neuronal pour classifier mes données textuelles, ce qui constitue une contribution méthodologique puisque l'analyse textuel en économie s'est limitée jusqu'ici principalement à des méthodes beaucoup plus simples avec d'importantes limitations. Je combine ces données avec les messages que les modérateurs du groupe de contre-discours se sont envoyés en interne qui contiennent notamment tous les articles que le groupe a considéré comme cible potentielle pour leurs interventions. Ceci me permet

d'identifier l'effet de traitement de ces interventions en comparant un groupe de traitement et un groupe témoin d'individus. Le groupe de traitement comprend tous les individus ayant commenté un article qui à été ultérieurement ciblée par une intervention. Le groupe témoin est constitué d'individus qui ont commenté des articles « concurrents » qui ont été considérés comme cible potentielle par les modérateurs du groupe de contre discours, mais qui n'ont pas été retenus. Le groupe se trouve souvent face à une contrainte de capacité : alors qu'il y a plusieurs articles qui méritent son intervention, il n'en choisit qu'un pour ne pas diviser ces forces. Je compare donc des articles qui ont un nombre comparable de commentaires, « likes » et messages haineux avant l'intervention afin d'assurer que la répartition des individus dans le groupe de traitement est exogène à leur comportement.

Le résultat principal de mon analyse est que les interventions de contre-discours ont un effet substantiel mais transitoire sur le comportement futur des individus. Pendant à peu près deux semaines après une intervention, les utilisateurs dans le groupe de traitement ont une probabilité plus faible de tenir des propos haineux que leurs homologues dans le groupe témoin. L'ampleur de cet effet est de 21%. L'effet émane principalement d'individus qui n'écrivent des messages haineux que sporadiquement et je ne trouve pas d'effet sur des utilisateurs qui se prononcent de cette manière plus d'une fois par semaine. De plus, les individus ciblés ont tendance à s'abstenir de débattre autour de sujets qui sont davantage susceptibles de déclencher des réactions xénophobes. Ils réduisent le nombre de commentaires qu'ils écrivent ou « aiment » de manière générale, et en particulier sur des sujets liés à l'immigration.

De plus, je démontre que les interventions de contre-discours ont un impact sur l'identité des individus qui participent au débat en ligne. D'abord, les interventions attirent de nouveaux utilisateurs n'ayant pas participé à l'intervention, augmentant ainsi le nombre de réactions aux articles ciblés de 50% au total. Ces individus sont plus modérés : ils ont une probabilité plus faible de se prononcer de manière haineuse et d'avoir un antécédent haineux. Par conséquent, alors que le nombre total de messages haineux sur les articles ciblés reste comparable au groupe témoin, leur part dans le nombre total de commentaires baisse de 3 points de pourcentage.

La propagation d'information ne semble pas pouvoir expliquer ces résultats. On pourrait imaginer que les messages écrits lors des interventions contiennent de nouvelles informations permettant aux individus ciblés de corriger des croyances a priori erronées et d'adapter leur comportement par la suite. Par exemple, les interventions pourraient amener des individus à reconnaitre que la criminalité parmi les refugiées est en fait plus faible qu'ils ne croyaient. Cependant, il n'y a qu'une petite partie des messages des contre-discours dans mon échantillon qui contiennent des références à des informations, ce qui suggère qu'il ne s'agit pas d'un traitement d'information. En plus, les effets sont à la fois temporaires et décroissant dans le nombre de traitements par individus, deux faits qui sont difficiles à réconcilier avec un effet de provision d'information.

Le mécanisme qui semble le plus cohérent avec mes résultats est que les interventions servent de sanctions non-pécuniaires qui communiquent la présence d'une norme sociale ou qui la font ressortir de manière plus saillante. Dans le contexte que j'étudie, l'efficacité des normes sociales

ne semble pas s'expliquer par le fait que les individus infèrent des normes du comportement moyen des autres utilisateurs et alignent leur comportement avec celui des autres. Par exemple, je ne trouve pas de corrélation entre la taille de l'intervention de contre-discours et son effet sur le comportement des individus. Il semble plutôt que le changement de leur comportement est déclenché par la désapprobation exprimée dans les messages de contre-discours – comme s'il s'agissait d'une sanction pour une transgression. Par exemple, l'effet des interventions est le plus fort pour les utilisateurs ayant reçu une réponse directe à leur commentaire de la part d'un membre du groupe de contre-discours au lieu d'un message qui condamne les discours haineux en général. Ce résultat est en accord notamment avec la littérature portant sur les sanctions non-monétaires, qui met en avant le rôle de ces sanctions pour communiquer des normes sociales.

Si nous souhaitons surtout éviter que les discours haineux puissent atteindre un large public, le fait que ces interventions réduisent la prévalence de messages haineux sur les pages Facebook de grands médias constitue déjà une bonne nouvelle. Cependant, on pourrait craindre que les individus ciblés par le contre-discours s'expriment tout simplement ailleurs, potentiellement dans des environnements qui pourraient favoriser une radicalisation plus forte. Bien évidement je n'observe les individus que s'ils commentent l'actualité en public sur Facebook, mais toujours est-il que les médias présents dans mon échantillon varient fortement en termes de prévalence de messages haineux et de probabilité d'être visé par des contre-discours. Je ne trouve aucun effet qui suggérerait que les individus migreraient systématiquement vers des médias qui attirent plus de xénophobes suite à une intervention. Il semble donc improbable qu'il y ait un fort effet de radicalisation qui amènerait ces gens à s'exprimer dans des environnements plus favorables à des propos extrêmes.

Ces résultats indiquent que les normes sociales sur les réseaux sociaux sont une arme à double tranchant. Alors que maintenir le dialogue malgré des différends peut aider à apaiser des débats parfois toxiques, le mécanisme décrit si dessus pourrait fonctionner aussi bien dans l'autre direction : des discours haineux pourraient entrainer plus de discours haineux. Dans ce contexte il est particulièrement inquiétant que les médias semblent avoir une incitation à produire du contenu provoquant des réactions haineux pour attirer l'attention sur les réseaux sociaux. Cela met en évidence une fois de plus la responsabilité importante dont doivent faire preuve les médias.

2e chapitre – Mesurer le goût pour l'image social

Comme je l'ai évoqué en introduction, la mesure dans laquelle les individus réagissent aux normes sociales est déterminée en partie par leur goût pour l'image sociale. Alors qu'ils varient en termes de motivation intrinsèque pour un acte prosocial, même un individu sans une telle motivation peut suivre une norme sociale si le fait de le faire lui confère suffisamment d'approbation ou, au contraire, si ne pas la suivre est suffisamment stigmatisant. Il existe une littérature empirique considérable démontrant que le goût pour l'image joue un rôle pour

expliquer le comportement des individus en moyenne. Cependant, nous savons très peu sur les déterminants et conséquences de ce goût pour la réputation au niveau individuel. Une des raisons principales pour cela est qu'il n'existait pas une mesure systématique et individuelle de la sensitivité du comportement à l'image sociale. Si des études le font, c'est toujours en utilisant des mesures très spécifiques au contexte étudié et donc difficiles à généraliser.

Le deuxième chapitre de ma dissertation est un article co-écrit avec Emeric Henry dans lequel nous présentons une expérience novatrice conçue pour mesurer le goût pour l'image au niveau individuel. L'expérience permet d'identifier ce paramètre indépendamment d'autres préférences sociales, dont notamment l'altruisme. Le jeu que nous proposons comprend trois joueurs : un dictateur, un récepteur, et un observateur. Le dictateur décide le montant de sa mise dans une loterie avec deux résultats possibles. Si le résultat est positif, le récepteur gagne une somme fixe d'argent. Dans le cas contraire, il ne reçoit rien. Plus la mise du dictateur est élevée, plus les chances de gain dans la loterie sont élevées. Le dictateur prend sa décision sachant que l'observateur sera informé du résultat de la loterie. Avant que le résultat de la loterie soit révélé, le dictateur doit indiquer combien il est prêt à payer pour pouvoir rester anonyme au cas où le résultat de la loterie est négatif, c'est-à-dire pour éviter que sa photo ne soit visible par l'observateur. Le récepteur ne voit jamais les photos des autres joueurs. L'observateur ne voit que le résultat de la loterie et ignore le montant que le dictateur y a investi au profit du récepteur.

La structure de ce jeu est motivée par deux aspects principaux. Premièrement, le goût pour l'image sociale du dictateur est directement mesuré par sa propension à payer pour rester anonyme au cas où le récepteur ne reçoit pas d'argent de la loterie. Deuxièmement, nous démontrons que sous des hypothèses raisonnables sur la fonction d'utilité du dictateur, cette mesure est indépendante d'autres préférences sociales, dont l'altruisme. Dans le cas où le dictateur ne reste pas anonyme, l'observateur n'est pas informé de la contribution à la loterie mais seulement de son résultat. De ce fait, l'inférence qu'il fait quand il voit la photo est une révision de sa croyance concernant les caractéristiques du dictateur conditionnée sur le fait que la loterie était un échec, mais il ne peut pas conditionner sa croyance sur la contribution à la loterie directement. Pouvoir séparer notre mesure d'autres préférences sociales est essentiel pour étudier les déterminants de l'image sociale et identifier d'autres caractéristiques auxquelles elle est corrélée.

Nous constatons une hétérogénéité considérable en termes de goût pour l'image parmi les participants d'une expérience en laboratoire. Un tiers des participants n'a aucune disposition à payer pour son anonymat. Alors que la plupart des participants offre des montants supérieurs à zéro, un autre tiers des participants offre des montants très importants. Peu de caractéristiques de l'observateur semblent affecter la propension à payer des dictateurs. Une exception notable est la nationalité : les observateurs étrangers offrent plus pour rester anonyme quand ils sont observés par un français. Une explication potentielle est que ces participants craignent que des préjugés pourraient amener les observateurs français à juger un résultat négatif de la loterie plus sévèrement.

D'autre part, le goût pour l'image sociale est-il lié à d'autres préférences sociales? Après avoir joué au jeu décrit ci-dessus, les participants ont procédé à un jeu de dictateur infiniment répété. Dans la moitié des séances expérimentales, le dilemme du prisonnier a été joué avec des observateurs, dans les autres sans observateurs. Dans les séances avec observateurs, nous ajoutons un troisième joueur avec la simple tâche d'observer le comportement des autres joueurs et d'y attribuer une note après chaque tour du jeu. Cela nous permet de documenter quel type de comportement est jugé favorablement par les participants et d'identifier la norme sociale prévalente.

Utilisant ces jeux répétés nous démontrons dans un premier temps que les individus avec un fort goût pour l'image ont tendance à moins coopérer dans le jeu quand ils ne sont pas observés. Nous interprétons ce résultat comme la preuve que ces individus sont plus égoïstes. Dans un deuxième temps, nous comparons les jeux joués sans et avec observateur pour montrer que les individus avec un fort goût pour l'image adaptent davantage leur comportement à la norme sociale que les autres – du moins lorsque leurs actes sont observables.

Ce chapitre introduit une méthode pour mesurer le goût pour l'image individuelle de manière systématique, chercher à valider cette mesure et commence à explorer ses déterminants et conséquences. Du fait de la nature du concept mesuré, l'expérience est moins simple à mettre en place que d'autres jeux, comme par exemple le jeu de confiance ou de dictateur. Toutefois les résultats de ce chapitre, notamment le fait que les caractéristiques de l'observateur importent peu, permettront de simplifier l'expérience selon le besoin tout en préservant les avantages clés de la structure du jeu. Par exemple, si les participants n'ont pas à craindre d'éventuelles représailles de la part des récepteurs, le récepteur peut servir d'observateur en même temps au lieu d'ajouter un troisième jouer. Ceci peut s'avérer utile par exemple pour travailler avec de grands échantillons en ligne. Dans des expériences de terrain, les participants pourraient tout simplement se lever de leur chaise pour révéler leur identité au lieu de prendre des photos. Nous avons cherché à tester la robustesse de l'expérience pour permettre à d'autres chercheurs de se baser directement sur nos résultats.

3e chapitre - Intégration verticale et verrouillage

Quand le producteur intègre verticalement l'un de ses acheteurs et arrête de fournir à ses concurrents en aval ayant besoin de ses produits ou services pour leur production, on parle de verrouillage vertical. Ceci peut être le cas par exemple quand l'entreprise en amont dispose de l'accès à une infrastructure ou technologie nécessaire pour la production du bien final et l'entreprise utilise ce fait pour empêcher d'autres firmes de lui faire concurrence sur le marché en aval. Les acheteurs rivaux sont ainsi verrouillés.

Il existe une importante littérature théorique examinant les motivations et conditions qui peuvent amener les entreprises à adopter cette stratégie. La théorie moderne, post-Chicago, de verrouillage de marché est basée sur deux papiers phares qui datent d'il y a presque trente ans : Hart and Tirole (1990) établissent des conditions sous lesquelles les entreprises peuvent

étendre leur pouvoir de marché en amont au marché en aval, tandis que Ordover et al. (1990) soulignent le verrouillage du marché comme un moyen d'augmenter le coût des intrants pour les entreprises concurrentes en aval. Cependant, sur le plan empirique, la question de savoir si les entreprises utilisent en pratique le verrouillage du marché se limite à un nombre restreint d'études de cas dans des contextes souvent très spécifiques. Cela s'explique par le fait que les relations verticales sont rarement observables par les chercheurs, limitant ainsi notre capacité à tester les théories. Nos connaissances sont encore plus limitées sur la façon dont les entreprises peuvent réagir aux menaces de verrouillage du marché afin d'en atténuer l'impact.

Par conséquent, le débat politique sur la régulation de l'intégration verticale par les autorités de la concurrence est loin d'être clos. En principe, de nombreuses juridictions considèrent le verrouillage vertical comme une violation du droit de la concurrence. Aux États-Unis, les tribunaux ont établi une doctrine sur le verrouillage de marché il y a plus d'un siècle. Pourtant, le niveau de rigueur que les autorités devraient appliquer pour veiller au respect de ces règles continue de faire l'objet de débats. Au moment de la rédaction de la présente dissertation, le ministère américain de la Justice et la Federal Trade Commission mettent à jour leurs lignes directrices sur les intégrations non-horizontales datant de 1984, et certains économistes et juristes réclament une politique de la concurrence "revigorée" (Baker et al. (2019)).

Dans le droit fil du débat en cours, les autorités de la concurrence ne contestent que rarement des fusions et acquisitions verticales. Dans le dernier chapitre, mon co-auteur Johannes Boehm et moi-même souhaitons déterminer si cela résulte d'une application laxiste des règles de la concurrence ou du fait que le verrouillage est une hypothèse principalement académique qui ne se produit pas en réalité. Quels sont les facteurs qui déterminent la prévalence du verrouillage vertical et quelle est l'ampleur des conséquences? Comment les entreprises menacées de verrouillage peuvent-elles en atténuer l'impact?

Pour répondre à ces questions, nous examinons empiriquement le verrouillage vertical dans une multitude d'industries et de pays. Pour ce faire, nous exploitons un nouvel ensemble de données de panel sur les relations verticales entre grandes entreprises, tant aux États-Unis qu'à l'étranger. Ces données nous permettent d'étudier si les relations acheteur-vendeur sont rompues plus fréquemment à la suite de fusions et d'acquisitions verticales. Nous montrons que les relations acheteur-vendeur sont plus susceptibles de se rompre lorsque le fournisseur s'intègre verticalement et que l'entité issue de la concentration en aval est en concurrence avec l'acheteur, mais pas lorsque l'entité intégré en aval n'est pas un concurrent de l'acheteur. Conformément aux théories du verrouillage de marché, la probabilité de rupture est encore plus élevée quand il y a peu de concurrence dans l'industrie en amont. Nous excluons que ces résultats s'expliquent par des chocs communs au niveau de l'industrie (ou au niveau des paires d'industries) à l'activité de fusion ou à la probabilité de rupture.

Bien entendu, cette corrélation n'implique pas en soi qu'il y ait du verrouillage vertical parmi les entreprises que nous étudions. Pour donner une interprétation causale à nos résultats, nous abordons trois menaces potentielles à leur validité. Premièrement, nous excluons la causalité inverse découlant, par exemple, d'un motif de sauvetage de fournisseurs en faillite. Les acheteurs

pourraient ainsi acquérir un producteur en amont pour assurer approvisionnement en intrants. Nous montrons que nos résultats persistent lorsque la probabilité d'intégration des fournisseurs est instrumentée par les flux de capitaux sortant des fonds communs de placement détenant leurs actions. Quand les fonds décroissent la taille de l'ensemble de leur portefeuille, ils exercent une pression à la baisse sur le cours des actions des entreprises qu'ils détenaient pour des raisons non liées à leur performance économique.

Deuxièmement, nous soutenons qu'il est peu probable que la corrélation soit le résultat de chocs communs. Nous répétons notre analyse sur l'échantillon de rumeurs sur les fusions et acquisition qui ne se sont jamais concrétisées et nous ne trouvons aucun impact. Dans la mesure où ces rumeurs d'intégration sont causées par les mêmes chocs non observés que les intégrations réelles, elles constituent un bon groupe de comparaison.

Enfin, la rupture des liens pourrait simplement résulter du fait que les parties intégrantes sont tellement plus efficaces que l'acheteur décide de quitter le marché. Cependant, nous montrons que les entreprises dont le concurrent s'intègre verticalement et qui n'ont pas de relation acheteur-fournisseur avec le fournisseur intégrant ne connaissent pas de baisse de chiffre d'affaires, ce qui suggère qu'il est peu probable que de fortes synergies puissent expliquer nos résultats.

Deux autres résultats complètent le dernier chapitre. Premièrement, les entreprises qui ont un motif de verrouillage sont plus susceptibles que les autres de fusionner avec un fournisseur. Il y a motif de verrouillage lorsque le fournisseur d'une entreprise vend également à son concurrent. Parmi les relations verticales actives, c'est précisément entre entreprises ayant un tel motif que l'intégration est la plus probable. Deuxièmement, nous étudions la performance des entreprises à la suite de l'intégration de leurs fournisseurs. Nous constatons une baisse marquée mais transitoire des chiffres d'affaires des entreprises dont le fournisseur s'intègre à l'un de leurs concurrents. La baisse des ventes est particulièrement prononcée pour les entreprises qui n'ont pas déjà un fournisseur de la même industrie que le fournisseur intégrant. Il s'agit là d'un moyen important pour les entreprises d'atténuer l'impact du verrouillage du marché : la diversification de la base de fournisseurs.

Nous interprétons nos résultats comme appuyant le point de vue selon lequel le verrouillage vertical du marché se produit dans la population des entreprises et des relations que nous étudions. Évidemment, des mises en garde s'imposent. Premièrement, comme nous ne disposons pas de données sur les prix et les quantités, nous ne pouvons pas nous prononcer sur les effets sur le bien-être des intégrations verticales que nous étudions. Nous nous concentrons plutôt sur le rôle de la structure du marché et sur la répercussion sur les entreprises concernées. Deuxièmement, notre échantillon se compose principalement d'entreprises qui sont cotées en bourse ou qui émettent des titres négociés et constitue donc un groupe sélectif. Toutefois, étant donné que ces entreprises et relations seront plus susceptibles d'être sous la loupe des autorités de la concurrence, nous pensons que le verrouillage vertical est probablement répandu également en dehors de l'échantillon que nous avons étudié.

References

Allcott, Hunt. 2011. "Social norms and energy conservation." *Journal of Public Economics* 95 (9-10): 1082–1095.

Baker, Jonathan B., Nancy L. Rose, Steven C. Salop, and Fiona M. Scott Morton. 2019. "Five Principles for Vertical Merger Enforcement Policy." *Georgetown Law Faculty Publications and Other Works,* no. 2148.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652–1678.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and so Should You." *Journal of Politics* 71 (01): 178.

Hart, Oliver, and Jean Tirole. 1990. "Vertical Integration and Market Foreclosure." *Brookings Papers on Economic Activity. Microeconomics* 1990 (1990): 205–286.

Ordover, Janusz A., Garth Saloner, and Steven C. Salop. 1990. "Equilibrium Vertical Foreclosure." *American Economic Review* 80 (1): 127–142.