Three Essays in Applied Economics

Konstantin Werner

TESI DOCTORAL UPF / Year 2022

THESIS SUPERVISOR
Albrecht Glitz
Department of Business and Economics, UPF

Thanks

Above all, I am deeply grateful to my advisor Albrecht Glitz. Only because of his guidance, endless patience, and his challenging but always helpful comments was I able to finish this dissertation. I would especially like to thank him for being understanding and supportive during difficult times; I consider myself fortunate that he was my mentor.

I also greatly benefited from many insightful remarks during seminars in the department and would like to thank Aina Gallego and Maria Petrova for fruitful discussions about my projects.

Further, I want to thank Marta and Laura to whom I could always turn for any questions. On many occasions, they helped me far more than I could have expected.

I also want to thank my friends and colleagues who have made the PhD a more pleasant and memorable experience. Ilja and Sébastien for being great office mates and friends, and for offering advice and helping me with their experience. Benny, Carla, Konstantin, Lars, Nils, Pascal, Paul and Steffen for always having an open ear, for motivation and support. Especially Steffen for many mind-clearing rounds of golf and Lars and Carla for helping me find my way after the PhD. Last but not least, Giulia and Rea whom I met during the last part of my PhD and who have made this period much more enjoyable. You being my flatmates made our flat a true home for me.

I feel deep gratitude for my parents. They always supported and motivated me in all my endeavors and shaped me to be the person who was able to finish this project. I am forever grateful to my father who, with his selfless help and infinite support during the most difficult times, gave me the strength to continue this dissertation. I hope that my mother would be happy and proud to see that I have successfully finished this chapter of my life.

Finally, I want to express my deepest gratitude to Julija. We met in Barcelona during my masters and you accompanied me for most of my journey through the PhD. You are the reason why I will always have a great memory of this time and this city.

Abstract

In this dissertation, first, I study the effects of broadband internet on individual and regional outcomes. Second, I provide an overview of the political espionage activities of the former East German secret service, the Stasi. In chapter one, I link the declining concentration of German vote share distributions to individual shifts from face-to-face towards virtual interactions. I show that regions where voters engage in more interactions over the internet exhibit less concentrated vote share distributions. In the second chapter, I investigate whether broadband internet influences the integration of immigrants in Germany. I find positive effects on employment probabilities, the probability to speak German at home and the frequency of performing voluntary activities. In chapter three, Albrecht Glitz and I uncover the thematic emphasis of the Stasi's political espionage. We further provide evidence that the Stasi systematically collected information on negotiations between high-ranking East and West German politicians.

Resumen

En esta disertación, primero, estudio los efectos de Internet en los resultados individuales y regionales. Segundo, describo el espionaje político de el servicio secreto de Alemania del Este, la Stasi. En el capítulo uno, vinculo la disminución de la concentración de las distribuciones de votos en Alemania con los cambios individuales de las interacciones cara a cara frente a las virtuales. Muestro que las regiones donde los votantes participan en más interacciones a través de Internet exhiben distribuciones de votos menos concentradas. En el segundo capítulo, investigo si Internet influye en la integración de los inmigrantes en Alemania. Encuentro efectos positivos sobre las probabilidades de empleo, de hablar alemán y la frecuencia de actividades de voluntariado. En el capítulo tres, Albrecht Glitz y yo descubrimos el énfasis temático del espionaje de la Stasi. Además, proporcionamos evidencia de que la Stasi recopiló información sistemáticamente sobre las negociaciones entre ambos estados alemanes.

Preface

In the first two chapters of this dissertation, I investigate how the shift from face-to-face towards virtual interactions, facilitated by the introduction of broadband internet, shapes individual and regional socio-economic and political outcomes. The basic intuition which underlies both chapters is the fact that individual decisions are prone to peer effects which lead to different outcomes for face-to-face and virtual interactions.

In chapter one, I show how the change in individuals' communication patterns can explain the decline of the federal vote share distribution that is observable for Germany since the mid-1970s. Intuitively, real-life social ties tend to emerge relatively more frequently between individuals with different political opinions than it is the case for virtual social ties. Therefore, peer effects in face-to-face interactions have the potential to align individuals' political opinions and thus concentrate the vote share distribution. Virtual social ties formed over the internet, on the other hand, are likely to emerge between voters of the same parties and thus cannot have any effect on vote share distributions.

I also build on the notion that research on mass media and peer effects suggests that the internet has significant effects on immigrant integration. The empirical studies investigating this relationship, however, focus only on a narrow set of integration dimensions and suffer from endogeneity. I try to fill these gaps with the paper presented in chapter two.

In the last chapter of this dissertation, Albrecht Glitz and I analyze a dataset on the political espionage activities of the so-called Stasi, the former East German secret service. On top of a comprehensive description, we also provide two examples of how to exploit the respective data for quantitative research, an approach that has not been implemented by many researchers in the field.

In chapter one, to show how virtual interactions affect aggregate political outcomes, I simulate a game with peer effects in voting for individuals placed on a network. Two parameters determine the concentration of the aggregate vote share distribution in the network. The overall number of social ties and the probability with which ties emerge between voters with similar partisan affiliation. Using this intuition, I link the decline in the concentration of German federal election vote shares since the mid-1970s to two key trends in the society. A decline of face-to-face and an increase of virtual interactions which take place relatively more often between individuals with similar partisan affiliation. To corroborate the assumptions of my model, for a sample of movers, I show that destination region vote shares significantly influence partisan affiliation. I provide evidence that the relationship is caused by peer effects and not by regional sorting. Using clubs and associations as a proxy for face-to-face interactions and the location of district courts in which they must be registered as an instrument, I confirm the model's

prediction that regions where voters have fewer social ties exhibit less concentrated vote share distributions. Finally, by implementing a well-established instrumental variable approach for the number of households in a region connected to the broadband network, I show that more virtual interactions further contribute to the decline in the concentration of the vote share distribution.

In chapter two, I study a panel of foreign-born individuals in Germany and assess whether broadband internet exposure influences a comprehensive set of integration indicators. To identify potential effects, I regress individual changes in integration outcomes on changes in individual and household level broadband exposure measures. Furthermore, I use the previously mentioned instrumental variable to predict technical broadband availability at the municipality level. Since the instrument is not sufficiently strong to implement a two-stage least squares regression, I opt for a placebo-type approach instead: although technically not available in their municipality of residence, some households still bought a broadband access for their homes. I exploit this fact and show that individual and household level measures of broadband exposure are positively related to migrants' employment probabilities, their probability to usually speak German at home and the frequency of performing voluntary activities. However, in line with a causal interpretation, these effects are only significant for individuals living in municipalities where broadband internet is technically available. They are not significant for those individuals who bought a broadband access but live in regions where broadband internet is technically not available. There is some evidence for heterogeneity. The effect of broadband exposure on employment possibilities decreases with a migrant's age. Effects on German language use and contacts to Germans become smaller when immigrants use the internet to communicate with friends and family from their countries of origin.

In the last chapter, together with Albrecht Glitz, I provide an overview of the political espionage activities of the East German secret service, the so-called Stasi. The basis of our analysis are meta data on politically relevant material that was sent to the Stasi by its spies in the West. These data also contain information about the reports that were compiled on the basis of the incoming materials and sent to high-ranking politicians and friendly intelligence agencies. We begin by giving a basic description of the form, relevance, and content of the respective materials. We then describe the spies and recipients in terms of the number and relevance of materials provided/received. In two analyses, we then show how to exploit the meta data, especially keywords used as content labels, for further research. First, we provide evidence that the Stasi targeted negotiations between high-ranking East and West German representatives on a cultural agreement and a significant loan. Second, we fit a Latent Dirichlet Allocation (LDA) model to uncover the thematic emphasis of the Stasi's political espionage and how it evolved over time.

Contents

List of Figures

List of Tables

Chapter 1

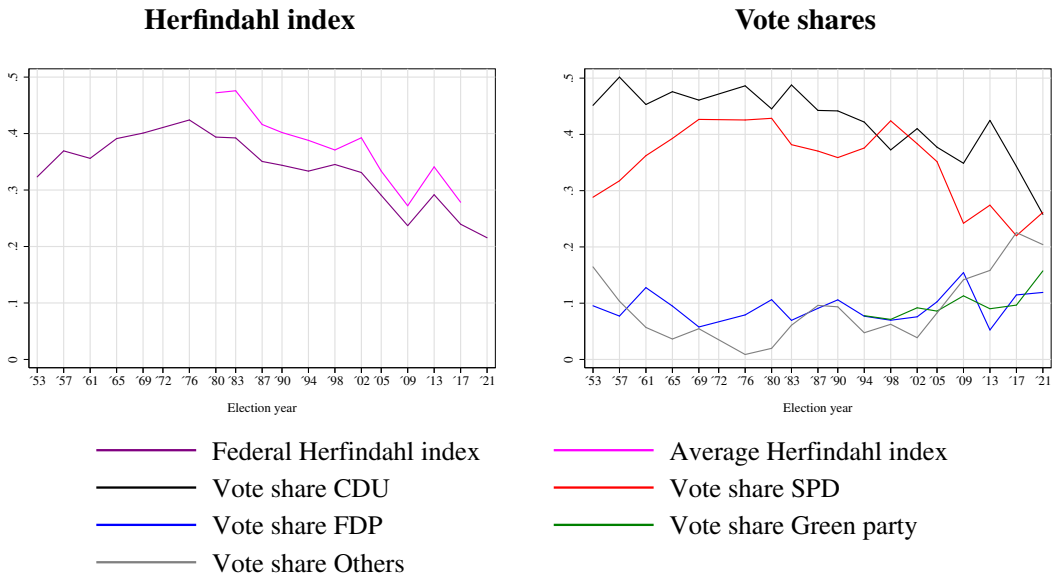
FACE-TO-FACE VERSUS VIRTUAL INTERACTIONS AND THEIR EFFECTS ON POLITICAL FRAGMENTATION IN GERMANY

1.1. Introduction

Since the mid 1970s, the distribution of vote shares over parties in German federal elections has become significantly more uniform. Figure 1.1 displays this trend for West Germany[1] by plotting the development of the Herfindahl index of vote shares from federal elections since 1953. For a given number of parties, the index is smallest if vote shares are uniformly distributed and highest if one party receives all votes. I will refer to the former situation as *dispersed* or *fragmented* and to the latter as *concentrated* or *polarized* vote share distribution. The purple line in the left panel of Figure 1.1 is based on election results aggregated to the federal level. The magenta colored line plots the municipality level average Herfindahl index (disaggregated data is only available from 1980 onward) to emphasize that the observable federal trend is not due to composition effects but that dispersion has indeed increased in the average municipality. Since 1980, the Herfindahl index decreased for 8228 of 8245 West German municipalities in my sample. Figure 1.A.1 in the appendix plots the respective empirical distribution. The right panel of Figure 1.1 shows how the increase in dispersion is related to shifts in vote shares between parties. Since the mid 1970s, the Christian-Conservatives (CDU) and the Social Democrats (SPD) lost vote shares to the Liberals (FDP), the Green party

[1] To avoid inconsistencies caused by the German reunification in 1990, I exclude East Germany from all analyses.

Figure 1.1: Trend of Herfindahl index and vote shares in postwar Germany



Notes: This figure plots the federal level Herfindahl index of vote shares in federal elections since 1953 and the corresponding non-weighted average municipality level Herfindahl index since 1980 in the left panel. The right panel plots the respective vote shares of major German parties. Election years are not evenly spaced because legislative periods can end before the regular election date. The graph of the average Herfindahl index begins in 1980 and ends in 2017 because municipality level voting data is not available for other years. Municipality level data on second votes for 1972 are missing.

(*B90/Die Grünen*) and other, former, low vote share parties.[2,3]

Since there has been no significant reform of the electoral system, the causes for this trend are unclear. In this paper, I link the increase in the dispersion of vote shares to two ongoing trends in the German society. First, a decline of face-to-face social interactions and, second, an increase in virtual social contacts between individuals, enabled by the introduction of broadband internet. Table 1.A.3 in the appendix provides an overview of these trends for representative samples of the West German population. To relate them to dispersion, I simulate a model where citizens' preferences over parties are represented by utility functions which are the sum of randomly distributed, idiosyncratic preference parameters and peer effects. Individuals assign higher utilities to parties when their peers vote for them. Observing their peers' past decisions, for example through discussions, and taking into account their own preferences, voters choose the party providing them with the highest utility. The model shows that vote share distributions are less concentrated for peer networks with fewer social ties or with ties very likely to form

[2]For a brief overview of the German electoral system and postwar federal elections, see Table 1.A.1 and Section 1.A.1 in the appendix.

[3]Table 1.A.2 in the appendix shows that the trend of dispersing opinions is also observable for other controversial topics in Germany.

between voters of the same parties, a characteristic of virtual tie formation.

In the empirical part of this paper, I pursue two goals. First, I corroborate the assumptions of my model and show that peer effects matter when individuals decide which party to vote for. Then I provide evidence for its two central predictions. Regions where individuals form fewer physical or more virtual social ties with each other, that is engage in less face-to-face or more virtual interactions, exhibit more dispersed vote share distributions.

Peer effects, or social influence, are significant factors in human interaction, especially in the context of political opinion formation. Subsequently, I will use both terms interchangeably and define them as the force that occurs when an individual's behaviors, opinions or emotions change (only) due to her interaction with other individuals or groups (Colman, 2015). Research has theorized about the importance of social influence effects (Granovetter, 1978; Banerjee, 1992) and confirmed their existence on many occasions (Conley and Udry, 2010; Funk, 2010; DellaVigna, List, and Malmendier, 2012; Banerjee et al., 2013; Bursztyn and Jensen, 2015; Chetty, Hendren, and Katz, 2016; DellaVigna et al., 2016; Perez-Truglia and Cruces, 2017; Perez-Truglia, 2018; Enikolopov et al., 2020; Bursztyn, Egorov, and Fiorin, 2020). There are several explanations for why individuals are prone to social influence but generally one can distinguish three overarching reasons (Kelman, 1958). First, individuals might believe that their peers have different or superior information than themselves and therefore change their attitudes and behaviours. Banerjee (1992) is a canonical example of this notion which Kelman (1958) refers to as *Internalization* and which is intuitively similar to *knowledge spillovers* in the literature on agglomeration effects (Glaeser et al., 1992). The second driver of peer effects is *Compliance* or social image concerns. Individuals care how they are perceived by others and therefore adjust their behaviour accordingly. Although image concerns have been shown to matter in the context of Political Economy (Funk, 2010; DellaVigna et al., 2016), they are unlikely to affect voting decisions which are not observable by peers. In the context of voting, it is more likely that peer effects either originate from Internalization or the third type of social influence, *Identification*. Identification occurs because an individual "wants to establish or maintain a satisfying self-defining relationship to another person or a group. [...] He adopts the induced behaviour because it is associated with the desired relationship" (Kelman, 1958). Advertisements featuring celebrities or, more recently, influencers rely on this type of peer effect.

To conceptualize aggregate consequences of peer effects in individual decision making, it has become a standard approach to use models of games played on networks. Within these frameworks, social influence can easily be incorporated as external ef-

fects through which individual decisions affect each others utilities (Jackson and Zenou, 2015). My model is an example of such a game featuring strategic complementarities in voting decisions (positive externalities/peer effects) and randomly distributed, intrinsic preferences over parties. To solve the game, I assume individuals behave myopically and choose parties as best response to the parties voted for by their peers.[4] In order to link the concentration of a region's vote share distribution to the number and type of social ties between its inhabitants, I explicitly model the tie formation process. Each period, individuals who are not peers have an exogenously determined chance to form a social tie. On the other hand, for individuals who are peers already, there is a possibility that ties between them will break. The probabilities of tie formation and breaking, in expectation, pin down a steady state number of overall social ties in the network. At the same time, utility maximization and myopic best response dynamics lead to a steady state vote share distribution which features one high vote share majority party and several low vote share minority parties. The wedge between the majority and minority parties' vote shares, and thereby the concentration of the vote share distribution, will depend on two key parameters. First, on the steady state number of social ties between individuals and second, on the probability with which these ties emerge between individuals who voted for different parties before becoming peers. In the literature on social networks, the second parameter is referred to as *direction* of the tie formation process.[5] Peer effects, incorporated as positive externalities in utility functions, increase overall utility when individuals vote for the same party, i.e. coordinate. Therefore, the vote share distribution has the tendency to tilt and produce one majority party. To understand the intuition of the model, it is useful to notice that it is similar to those used for studying product or service adoption in the presence of network effects (compare, for example, chapter 17 of Easley, Kleinberg, et al., 2010 or McGee and Sammut-Bonnici, 2015). Parties can be thought of as mutually exclusive products which provide higher utility to an individual if they are also "used" by her peers. Classical examples of such

[4]Using the concept of Nash Equilibrium to solve the model makes simulation more difficult. Furthermore, myopic individuals are more realistic in the context of (political) opinion formation. First individuals learn the opinions of their peers, for example during discussions. Only afterwards there is scope for social influence. Letting individuals form beliefs about the opinion of their peers and decide accordingly, would imply that social influence takes place without mutual knowledge of each other's opinions.

[5]It is well documented that individuals sharing social ties are similar in many dimensions, for example in their education levels, believes and values (McPherson, Smith-Lovin, and Cook, 2001). One explanation for this observation is the directed nature of the tie formation process. Individuals are more likely to form ties with like-minded others. This "love for the like-minded" is known as *homophily* and can either emerge due to preferences (individuals are drawn to similar others) or due to the fact that like-minded others are more likely to meet because they live in the same areas, have similar jobs or hobbies. Still, it is very unlikely that individuals who are not similar have zero chances of meeting. The opposite of a very directed process is referred to as *random* tie formation and, in the context of this paper, would imply that individuals form ties completely independent of the parties they are voting for.

products are messenger apps, online social networks or computer operating systems. In these models, market shares (here: vote shares) tend to tilt towards one good (political party), leading to a winner-takes-it-all scenario. A natural monopoly stabilized by network externalities (peer effects) emerges. Analogously, in my model, vote shares will tilt, such that a "monopolist" political party emerges.[6]

If the system moves from a steady state with many social ties to one with fewer ones, like it is the case for Germany as indicated by the decline in face-to-face interactions, the vote share distribution disperses. Alternatively, if the direction of the tie formation process increases, it is less likely that voters of different parties become peers, there is less scope for coordination and the wedge between majority and minority parties will be smaller. The increase of virtual interactions in Germany, facilitated by the introduction of broadband technology around the year 2000, was likely prone to a very directed tie formation process. A lot of research has focused on the directed nature of virtual tie formation and its role in the emergence of *Echo-Chambers*, peer networks in which individuals are confronted almost exclusively with opinions similar to their own. For example, Sunstein (2001) argues, while individuals are exposed to a variety of viewpoints in real-life, since some real-life social interactions occur randomly, these exposure ceases to exist if many aspects of physical life can be carried out online. Especially in the context of political opinions, virtual tie formation is found to be very directed (Adamic and Glance, 2005; Lawrence, Sides, and Farrell, 2010). More recently, the introduction of online social networks has created platforms like Facebook and Twitter which are explicitly dedicated to virtual tie formation and also frequently used to exchange political viewpoints. It has been confirmed that, indeed, tie formation on these platform is very directed with respect to political opinions (Conover et al., 2012; Barberá et al., 2015; Mosleh et al., 2021). Anecdotal evidence further implies that algorithms of internet giants like Amazon, Facebook, Google or Spotify seem to actively foster directed tie formation or media consumption (also a source of peer effects) since their suggestions of new content is heavily based on similarity to the contents individuals already consume (Dewey, 2015).

Of course, there are other possible explanations for the dispersion of the vote share distribution in Germany. It is likely that voters' preferences have changed over time or that former niche parties broadened the spectrum of their agendas. For example, the Green party has been winning votes (thereby dispersing the vote share distribution) be-

[6]Figure 1.1 showed that in Germany there was not one, but two majority parties. In the context of peer effects/network effects, this could be explained by the fact that political parties are not mutually exclusive products. A CDU voter might get additional utility also from his SPD voter contacts and vice versa. The actual exclusivity occurs between mainstream centrist parties on the one hand versus smaller niche parties on the other.

cause, on the one hand, individuals value environmental protection more than they used to. On the other hand, the Green party managed to position itself as economically competent, attracting new voters who had previously only associated it with environmental protection. Still, in the main part of my empirical analysis, I show that the decline in face-to-face and the increase in virtual interactions can at least explain parts of the dispersion in the vote share distribution.

I begin the empirical part of this paper by providing evidence that peer effects are indeed important for voting decisions. To do so, I follow an approach by Perez-Truglia (2018) and merge individual level panel data on residential locations, partisan affiliation and support with federal election results on the municipality level. I restrict my sample to individuals who changed residence and use federal election results in destination regions as proxy for the political opinion of individuals' peers. In line with the predictions of my model, post-move support for parties constituting the majority in the new municipality of residence is significantly higher than for minority parties. The effect size equals roughly 55% of the standard deviation of the partisan support variable across individuals. To support my claim that this correlation is caused by peer effects, I rule out possible alternative explanations like regional sorting. In my regressions, I include a set of fixed effects and further control variables conditional on which the allocation of movers into destination regions should be quasi-random. I corroborate this assumption by showing that, conditional on the included covariates, individual pre-move party support is not correlated with the respective party's majority status in the destination region. If movers were to sort into destination regions along political characteristics, this relationship should be positive. To provide further evidence for peer effects driving my results, I exploit the fact that the relationship between vote shares and party support should be stronger when individuals have more peers in their destination region. Therefore, I interact a party's majority status with the number of clubs and associations in the respective destination municipality. Clubs and associations facilitate face-to-face interactions in a region and make it easier for individuals to form social ties with each other. Many other authors in the literature use this variable to measure the number of social ties in a region (Putnam, 1993; Knack and Keefer, 1997; Rupasingha, Goetz, and Freshwater, 2006; Kesler and Bloemraad, 2010; Satyanath, Voigtländer, and Voth, 2017). Especially the index of Rupasingha, Goetz, and Freshwater (2006) has been widely applied.[7] Indeed, for a standard deviation increase in the number of clubs and associations above the average of 890 clubs per municipality, the effect of a party's majority status on partisan support increases by ca. 8% of a standard deviation.

The main part of the empirical analysis focuses on testing the model's key predic-

[7] A more detailed discussion of that variable is deferred to Section 1.4.

tions, that is the effects of face-to-face and virtual interactions on the German vote share distribution. Using the fact that clubs and associations have to be registered in a special district court, the so called *Amtsgericht*, I construct a *near-college*-type instrumental variable (Card, 1995), indicating the presence of a district court in a municipality. Results of the respective two-stage least square regression confirm that fewer physical social ties within a region are related to higher levels of dispersion. A standard deviation decrease in the number of clubs and associations causes the 2017 Herfindahl index of a municipality to decrease by about 16% of a standard deviation.

To assess the effects of virtual interactions, I use municipality level data on the share of households connected to the broadband network, the broadband penetration rate. It is a good proxy for the number of virtual social ties because accessing online social networks or connecting with other individuals over the internet is effectively only possible with a broadband internet connection. Since this variable is likely endogenous, I apply a first-difference approach and also use an instrumental variable constructed by Falck, Gold, and Heblich (2014). It exploits particularities of the German telephone network which had unintentional but strong and heterogeneous effects on broadband availability across German municipalities. The two-stage least square coefficient implies that greater increases in broadband penetration were related to greater increases in dispersion of regional vote shares. However, the effects are small: a standard deviation increase in broadband penetration implies only a 5% of a standard deviation increase of the (negative) difference in the Herfindahl index between 2002 and 2009. Finally, I provide evidence for the exclusion restriction by estimating a placebo regression. The coefficient of the instrument is insignificant in a regression explaining changes of the Herfindahl index before broadband was introduced in Germany in the year 2000.

The remainder of this paper is structured as follows: Section 1.2 gives an overview of the related literature. Section 1.3 formalizes my model and derives empirically testable predictions. Section 1.4 describes the data used throughout this paper. Section 1.5 details the empirical strategy to test for peer effects between voters and presents the respective results. Sections 1.6 and 1.7 are devoted to assess the predicted outcomes of the decrease in face-to-face and the increase in virtual interactions, respectively. Section 1.8 concludes.

1.2. Literature Review

This paper contributes to the literature on the potential causes of political fragmentation and polarization. Many researchers have identified the internet as a key driver for these phenomena because, through online blogs and social networks, it tends to con-

nect mainly like-minded individuals and thereby create Echo-Chambers. This view has been verbalized prominently by Sunstein (2001): "People restrict themselves to their own points of view — liberals watching and reading mostly or only liberals; moderates, moderates; conservatives, conservatives; Neo-Nazis, Neo-Nazis." For example, Adamic and Glance (2005) show that in the network of links which connected 1000 of the most prominent political blogs during the 2004 US Presidential election, liberals and conservatives were mostly connected within separate communities. Furthermore, according to Lawrence, Sides, and Farrell (2010), online media consumers are also more likely to visit those blogs which align with their political leanings. More recently, the introduction of online social networks has created platforms which are dedicated to virtual tie formation. Conover et al. (2012) show that Twitter users are significantly more likely to retweet messages similar to their own political positions. Similarly, Barberá et al. (2015) find that almost 75% of retweets on political topics are between users with comparable ideologies. Especially suggestive is the paper of Mosleh et al. (2021) who show that individuals are close to three times more likely to follow individuals whose partisan affiliation matches their own. The implicit assumption underlying the logic of these papers is that peer effects matter for political opinion formation and the lack of exposure to differing opinions cements ideological views and fosters fragmentation (Hargittai, Gallo, and Kane, 2008; Gaines and Mondak, 2009; Yardi and Boyd, 2010; Conover et al., 2011; Aragón et al., 2013; Colleoni, Rozza, and Arvidsson, 2014; Gruzd and Roy, 2014; Barberá, 2015; Garcia et al., 2015). However, this literature is concerned with fragmentation and polarization in the virtual sphere and not with potential effects on real-world phenomena like election outcomes.

Another strand of literature explores the consequences of peer effects in face-to-face interactions and its effects for spatial clustering of politically like-minded individuals, that is regional polarization. A theoretical example of this research is Latané (1996) who develops a framework that explains how social influence between individuals leads to similar "attitudes, values, practices, identities and meanings" within regions. Many empirical papers have used data on campaign contributions to infer and explain spatial patterns in political attitudes (Tam Cho, 2003; Gimpel, Lee, and Kaminski, 2006; Perez-Truglia and Cruces, 2017; Perez-Truglia, 2018). For example, in Perez-Truglia (2018), political conformity leads individuals to contribute more to Barack Obama's 2008 presidential campaign in areas inhabited by more Democrats.

Using the intuition of both the literature on online and face-to-face interactions, I develop a model to explain how the ongoing dispersion of the vote share distribution in Germany can be linked to the decrease in physical and the increase in virtual social ties. While it has been noted that the theory of games played on networks can be used

to model regional polarization (Valente, 2005; Van Alstyne and Brynjolfsson, 2005), explicitly incorporating the decline in face-to-face interactions and an increase in the direction of newly formed ties in one comprehensive framework has, to my knowledge, not yet been done. Furthermore, empirically assessing the effects of these trends on election outcomes also represents a novel contribution.

This work is also related to the literature on social capital because many definitions of the term include the number and/or strength of social ties between individuals within a region (Durlauf and Fafchamps, 2005). The decline in face-to-face interactions could be framed as a decline in social capital, a phenomenon extensively described by Putnam (2000). On the other hand, authors believe that virtual ties or interactions do not benefit social capital. For example, Olken (2009) argues that increased TV consumption has decreased social interactions, an effect that can likely be caused by the internet as well, especially if used passively. Franzen (2003) further points out that many activities previously related to real-life interactions, like shopping and banking, can now be carried out online without the necessity for human communication.

In testing whether social influence affects voting decisions, this paper also augments the vast literature on peer effects. In political contexts, for example, image concerns have been shown to affect the expression of xenophobic views (Bursztyn, Egorov, and Fiorin, 2020), the likelihood for partisan donations (Perez-Truglia and Cruces, 2017), or voter turnout (Funk, 2010; DellaVigna et al., 2016). Of course, there are far more contexts in which they are important, too, one example being performance amongst pupils, as shown by Bursztyn and Jensen (2015). A very interesting relationship is uncovered by Acemoglu, Reed, and Robinson (2014) who show that social capital (including measures of the number of physical social ties between people) in Sierra Leone is connected to lower levels of political competition (fewer ruling chieftons per region). Here, similar to my model and those concerned with spatial polarization, more face-to-face interactions and the implied peer effects lead to a more concentrated (or polarized) political spectrum. Surprisingly, the causality of this relationship seems to stem from the fact that strong leaders actively engage in the creation of social capital to secure their power. The potential of social interactions as means to preserve (political) power has also been noted by Coleman (1998) who points out that social image concerns deter individuals from behavior deemed undesirable by society and can therefore be (ab-)used as control mechanism.[8]

Finally, in their paper, Satyanath, Voigtländer, and Voth (2017) show that German

[8]Here, again, the similarity to models of natural monopolies caused by network externalities is striking. Just like peer effects deter individuals from buying products with low market shares, they can prevent them from adopting socially unwanted behaviors or, given the topic of this paper, electing minority parties.

regions with tighter and more social ties (measured by the per capita number of clubs and associations) exhibited higher support for the Nazi party before the second world war. The authors' explain that Nazi officials found it easier to exert social influence on potential voters in regions with more social ties. Besides serving as further reference for the suitability of the clubs and associations variable in my analysis, these results provide evidence that social influence in the context of political opinion formation can also stem from Identification. This finding is important because, as mentioned before, social image concerns cannot be used to justify peer effects in voting.

1.3. Model

In this section, I describe my model and explain how it captures decreases in face-to-face and increases in virtual interactions. By way of simulation, I show how these changes in social interactions lead to a dispersion of the vote share distribution. For each simulation result, I briefly outline my empirical strategy to test the respective predictions in the data.

Consider a representative region, inhabited by N individuals. Call the (finite) set of individuals I (with cardinality $|I| = N$) and assume that each individual $i \in I$ has preferences over the finite set of all political opinions O. The different political opinions in O are indexed by k, such that $o_k \in O$, $k = A, B, \ldots$.[9] The preferences are represented by s_{ik}, each of which are independently and identically distributed continuous random variables following a well-behaved probability distribution $f(s_{ik})$, defined on the entire real line.[10] When deciding which party to vote for at time $t \in T$[11], individuals take into account their preferences and the (observable) decisions of their peers. Let G_t be an NxN matrix, determining the relationships of individuals within a region at time t. For example, $g_{ijt} \in \{0, 1\}$, the element of G_t found in row i and column j, takes the value one if individuals i and j are connected by a social tie in period t and zero if not. I will refer to i and j as peers if $g_{ijt} = 1$. The network of social relationships is symmetric so

[9] I also assume that there is a different party for each opinion and that each opinion directly translates into a vote cast for the respective party. "Holding an opinion" and "voting for a party" mean the same.

[10] One does not need to interpret s_{ik} as the intrinsic preferences for a political opinion. It is enough to assume that these variables capture all aspects of the opinion formation process which are unrelated to social influence effects. This is especially important because there is no conceptual difference between an opinion and the preference for a party. In a model studying the effects of social image effects, one could distinguish between preferences/opinions on the one hand and expressed opinions/preferences on the other. For simplicity, I will still use the term preference.

[11] The model is discrete, so that T is a subset of the set of natural numbers $\mathbb{N}$.

that $g_{ijt} = 1$ implies $g_{jit} = 1$. By convention, $g_{iit} = 0$. Utility can then be written as:

$$U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\} \tag{1.1}$$

$o_{it} \in O$ represents the party individual i votes for in period t, $\mathbb{1}\{\cdot\}$ is the indicator function which equals unity if the condition in the brackets is met and zero otherwise. $\sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik}$ is the idiosyncratic part of utility. For example, if individual i chooses opinion $o_k = o_A$, her utility, in the absence of peer effects, would simply be equal to s_{iA} which is the realization of the random variable s_{ik}. $\delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$ represents the social influence effect. The utility an individual derives from holding an opinion increases proportionally to the number of her peers holding the same opinion. The strength of peer effects is measured by the parameter δ.[12]

Since an individual's utility is affected positively by the decision of her peers, this model can be viewed as a game of pure complements with constricted action space, played on a network. Most of the following considerations are based on Bramoullé and Kranton (2016) who derive many results for games of pure complements played on a network. However, their analysis of constricted action spaces focuses either on constricted continuous action spaces or binary action spaces (and not on multinomial action spaces, as considered here). For different realizations of preferences and different networks, different Nash Equilibria are possible. For example, if $|O| = 2$, that is there exist only two parties, Bramoullé and Kranton (2016) show an equilibrium exists for the utility function specified above. Using their intuition, there are two conditions that allow for the characterizations of equilibria for any network. First, the opinion distribution resulting from the preference distribution will be an equilibrium if:

$$\delta < \min_i \left[\frac{|s_{iA} - s_{iB}|}{d_{it}} \right] \tag{1.2}$$

where d_{it} is the degree of individual i at time t, that is the number of her peers. To understand condition (1.2), consider an individual who would elect party o_A in autarky, i.e. $\max_k [s_{ik}] = s_{iA}$, but has peers who exclusively hold opinion o_B. Then, according to the above utility specification, the individual would switch to opinion o_B if $\delta d_{it} + s_{iB} > s_{iA}$ or $\delta > \frac{s_{iA} - s_{iB}}{d_{it}}$. Condition (1.2) rules out this case: δ is so small that even if all the

[12] Figure 1.B.6 in the appendix shows qualitatively similar plots to the ones in this section for a specification with decreasing marginal utility in the peer effect, $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \sqrt{\delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}}$

individual's peers held o_B, $U_{it}(o_B) < U_{it}(o_A)$. This is a trivial result as it is equivalent to $\delta = 0$, i.e. the absence of social influence effects. Consider instead the case where

$$\delta > \max_i \left[\frac{|s_{iA} - s_{iB}|}{d_{it}} \right] \tag{1.3}$$

Further assume that G_t is path connected.[13] Now, each individual, even the one with the highest difference between her preference parameters, will vote for a party if all her peers vote for that party. This condition immediately implies that full coordination (on any party) is a Nash Equilibrium. The equivalent conditions for the multi party case are:

$$\delta < \min_{i,k,l} \left[\frac{|s_{ik} - s_{il}|}{d_{it}} \right] \tag{1.4}$$

$$\delta > \max_{i,k,l} \left[\frac{|s_{ik} - s_{il}|}{d_{it}} \right] \tag{1.5}$$

However, full coordination is not the only possible equilibrium. It is easy to construct a network with δ fulfilling (1.5) such that, still, different opinions can co-exist in equilibrium. In general, the set of possible equilibria depends on the realized distribution of preferences and on the network structure. One problem is then to determine which equilibrium is chosen by utility maximizing agents. Therefore, I assume that individuals, instead of forming beliefs about the decisions of their peers, behave myopically and condition their best responses on the past decisions of their peers. As mentioned in Footnote 4, in the context of (political) opinion formation this assumption is also more realistic than using the concept of Nash Equilibrium. Myopic, deterministic best response behaviour generates dynamics as described for example in Blume et al. (1993) or Young (2020). Still, for any given network and preference distribution, it is hard to determine if the opinion distribution will converge towards a Nash Equilibrium and if so, towards which one. Therefore, I simulate the model in the following way: assume that "at the beginning of time" (at $t = 0$) none of the individuals residing in a region are connected with each other. Each period, among all elements of G_t with $g_{ijt} = 0$, a set $\mathcal{F}_t$ with cardinality $|\mathcal{F}_t| = F_t$ is drawn randomly and all its elements are set to one, so that F_t new ties have formed. Likewise, each period, among all entries of G_t with $g_{ijt} = 1$, a set $\mathcal{B}_t$ with cardinality $|\mathcal{B}_t| = B_t$ is drawn randomly and its elements set to zero, so that B_t ties have broken. I set $F_t = c_F(N^2 - \sum_{i \in I} d_{it} - N)$, that is the number of new social

[13] For any individuals $i, j \in I$, there is a path (of arbitrary length) of social ties connecting these individuals. This assumption is without loss of generality because a network that is not fully path connected can, obviously, be described as the union of its path connected elements.

ties equals the product of the constant share c_F and the current number of zero entries in G_t. Analogously, $B_t = c_B(\sum_{i \in I} d_{it})$, so that each period a constant share c_B of existing social ties are broken. Define the total number of social ties in the network $\sum_{i \in I} d_{it}$ as D_t and note that the condition $F_t = B_t$ pins down its steady state value D^*:[14]

$$\begin{aligned} F_t &= B_t \\ D^* &= \frac{N(N-1)c_F}{c_F + c_B} \end{aligned} \tag{1.6}$$

Additionally, the parameter R regulates, for $\mathcal{F}_t$ and $\mathcal{B}_t$ alike, how much more (less) likely it is that zero entries (one entries) of G_t between individuals with the same opinion are drawn as compared to those between individuals with differing views. So R determines how directed the tie formation and tie breaking processes are. For example, if $R = 5$, it will be five times more likely that a social tie forms between individuals with the same opinion than between individuals with different opinions. Analogously, it will be five times less likely that a social ties breaks between individuals with the same opinion than between individuals with different opinions (since directed tie breaking implies that ties between individuals with opposing views break more frequently). A value of $R = 1$ implies that tie formation and breaking are random.[15] Finally, each period, individuals vote for the party which maximizes their utility, i.e. taking into account idiosyncratic preferences and the (past) opinions of peers.

Figure 1.2 shows results from simulating the model for $T = 300$ periods and $N = 100$ individuals. Each panel plots average trajectories across 50 simulation loops of vote shares and the Herfindahl index for different parameter values. Utility is specified according to equation (1.1), preference parameters s_{ik} follow standard normal distributions which are independent across individuals and parties. For the baseline specification in panel (a), $\delta = 0.2$, $c_F = 0.01$, $c_B = 0.05$ and $R = 1.5$. The black line corresponds to the vote share of the majority party of period $t = 0$. The red line to the vote share of the party having the second highest number of supporters in $t = 0$. The other lines are defined accordingly. The dashed purple line represents the Herfindahl index. Given the randomness of the preference distribution and tie formation, the plots visualize simulations of the *expected* vote share distribution over the majority, second strongest, third strongest and minority party. Note that, because all s_{ik} are identically

[14] The condition also pins down the average, individual steady state degree $d^* = \frac{D^*}{N}$.

[15] I assume that the ability to form social ties is independent of party preferences. A Social Democrat, for example, is not, generally, more likely to form social ties than a Conservative. Also, the probability of tie formation depends only on whether individuals hold the same or different opinions but not on the opinions themselves. For example, a Social Democrat is equally likely to form a tie with a supporter of the Green or the Conservative party.

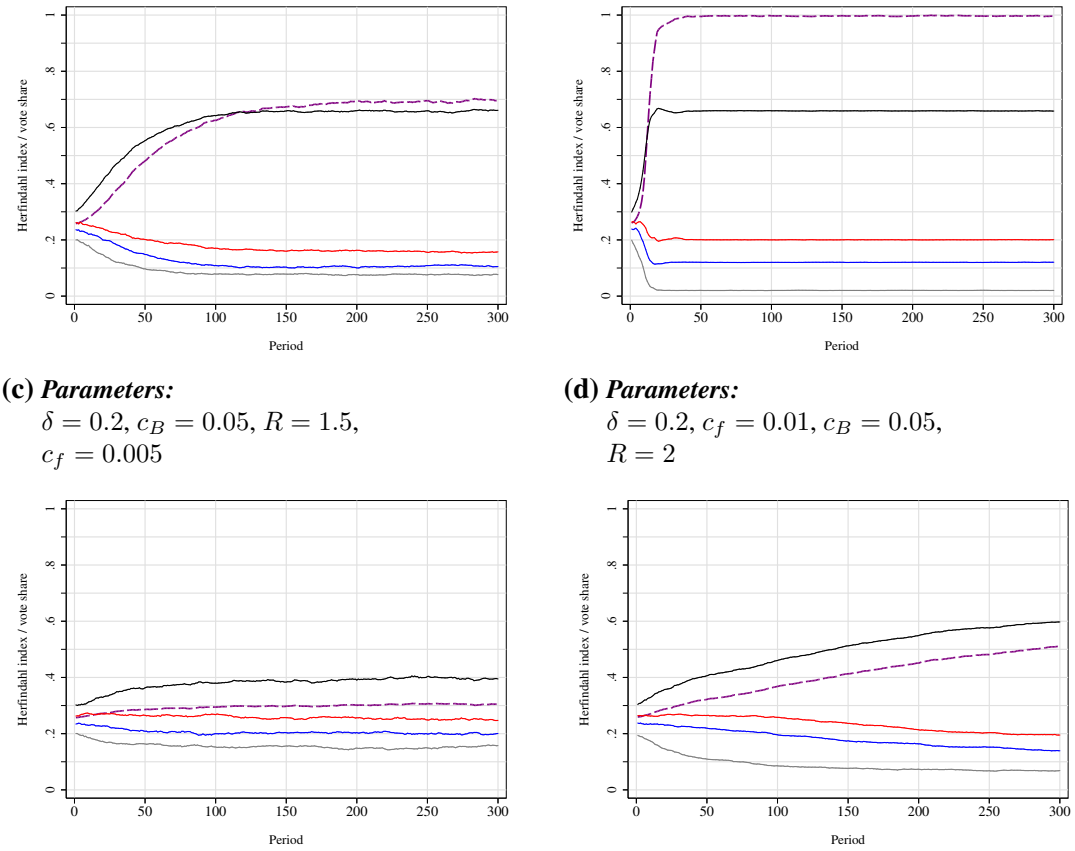
Figure 1.2: Simulation results for Herfindahl index and vote shares

(a) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 1.5$

(b) ***Parameters:***
$c_f = 0.01$, $c_B = 0.05$, $R = 1.5$, $\delta = 0.5$

(c) ***Parameters:***
$\delta = 0.2$, $c_B = 0.05$, $R = 1.5$, $c_f = 0.005$

(d) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 2$



Notes: This figure shows the average trajectories of vote shares and the Herfindahl index across 50 simulation loops of the model presented in Section 1.3. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals and $c_B = 0.05$. In each panel, one of the parameters δ, c_F and R is varied as compared to the baseline specification in panel (a).

distributed, in each simulation loop all parties have an equal probability (of $\frac{1}{|O|}$) of becoming the majority party in $t = 0$. Therefore, the initial majority parties across simulation loops are not necessarily the same. In other words, the respective plots in Figure 1.2 are simulations of the expected vote share distribution over parties, *conditional* on

the knowledge of the initial ranking of these parties.[16]

In period $t = 0$, vote shares are relatively similar, implying a dispersed political spectrum. As soon as individuals begin to form ties, the vote share distribution tilts towards the initial majority party (in expectation) and the Herfindahl index increases. Figure 1.B.1 in the appendix visualizes tie formation by plotting the average trajectories of the network-wide average degrees, d^*, for the respective simulations of Figure 1.2.[17] At some point, the process reaches a steady state and vote shares, the Herfindahl index as well as degrees remain constant.

To understand the intuition behind these trajectories, consider the baseline specification of panel (a). In $t = 0$, $d_{i0} = 0 \quad \forall \quad i \in I$. Then, since $c_F = 0.01$, $N = 100$ and $F_0 = c_F(N^2 - \sum_{i \in I} d_{i0} - N)$, 99 ties will emerge between the individuals placed in the network G_0. Imagine, for a moment, that tie formation (and breaking) ends now. For those voters who became peers and did not hold the same opinion prior to forming a tie, there is scope for coordination on the same party. If social ties created only dyads (two connected individuals), due to the identical distribution of preference parameters, it is entirely random whether any of the connected individuals switches opinions and if so, which one. However, if more than two individuals form a path-connected subnetwork, by definition, most of these individuals will probably hold the majority opinion. Then, across all possible network formations, in expectation, there are more individuals holding the majority opinion exerting a peer effect on minority opinion voters than vice versa. Therefore, it is most likely that majority opinion individuals "convert" minority opinion voters in these subnetworks. After a first iteration of utility maximization, the majority party has increased its vote share. In the next iteration, these dynamics repeat and even minority opinion individuals with high preference parameters might be converted because now more of their peers hold the majority opinion, exerting even stronger peer effects. A cascade emerges which leads to the winner-takes-it-all scenario characteristic for games with network effects. At some point, further iterations do not change the opinion distribution anymore. Either because all individuals are converted or because some individuals have such high preference parameters for the minority opinion that, given the network structure, peer effects are not strong enough to convert them. Now let the tie formation (breaking) process continue. With $c_B = 0.05$ and $B_t = c_B(\sum_{i \in I} d_{it})$, 5 ties will break (in my simulations, I round values to the closest

[16] A plot of the expected vote share distribution over parties, say o_A, o_B, o_C and o_D, instead of plots over initial majority party, second strongest party ... minority party, would show four relatively smooth lines at 25% since all parties are equal by definition. Defining parties according to their vote share ranking in $t = 0$ removes this equality.

[17] Figure 1.B.2 in the appendix additionally shows exemplary degree trajectories based on only one simulation and the same parameters.

integer). Now, since fewer individuals are connected with each other, the opinion distribution, in expectation, moves towards a more dispersed state (potentially in multiple iterations). When tie formation continues, $F_1 = c_F(N^2 - \sum_{i \in I} d_{i1} - N) \approx 98$ new ties will emerge in period $t = 1$. Even more individuals are peers now, further increasing the scope for a concentration of the vote share distribution. While individuals constantly form and break ties with each other, at some point a steady state is reached such that the number of new ties equals the number of broken ties. Until that steady state is reached, the expected concentration of the vote share distribution increases because more individuals can coordinate on the same party. However, in the steady state, while vote shares might still change, in expectation, the vote share distribution remains constant. Note that, in my model, new ties are formed and broken in each period, overlaying the iterative process of utility maximization. Still, the general dynamic of cascading or tipping leads the vote share distribution to become more concentrated, producing one majority and several minority opinions.

Simulating the average over many trajectories masks some of the dynamics. For example, due to the randomness of preferences and network structure, it is not always the initial majority opinion which profits from the winner-takes-it-all scenario. Figure 1.B.3 in the appendix plots some exemplary trajectories of the averages shown in Figure 1.2. The vote share distribution does indeed always tilt, however not always towards the initial majority party. To illustrate a similar problem, consider the differences between plots of panels (a) and (b) of Figure 1.2 caused by increasing the strength of the peer effect δ from 0.2 to 0.5. Despite this difference, the expected steady state vote shares of parties seem not to change too much. Still, the expected Herfindahl index is significantly higher when $\delta = 0.5$ (and steady state is reached faster). Figure 1.B.3 (b) reveals that the system is very close to the full coordination equilibrium in each simulation loop for $\delta = 0.5$. However, since individuals do not always coordinate on the initial majority party, its expected vote share is not one. Generally, however, the random element of tie formation "favors" individuals holding the initial majority opinion.

Panel (c) shows that the steady state Herfindahl index is significantly lower when c_F is smaller than in the baseline specification. c_F regulates how many new ties, F_t, are formed each period by determining the cardinality of the set $\mathcal{F}_t$ through the equation $F_t = c_F(N^2 - \sum_{i \in I} d_{it} - N)$. For smaller values of c_F, the steady state number of degrees D^* and the network-wide average steady state degree $d^* = \frac{D^*}{N}$ are smaller. Panel (c) in Figure 1.B.1 in the appendix visualizes this relationship. Analytically, this can be seen by taking the derivative of Equation (1.6) with respect to c_F and noting that it is

positive:

$$\frac{\partial D^*}{\partial c_F} = \frac{\partial \frac{N(N-1)c_F}{c_F+c_B}}{\partial c_F} > 0 \tag{1.7}$$

When degrees are lower, fewer voters are peers so that the scope for coordination on one party decreases and therefore dispersion is higher. Equation (1.6) also shows that increasing c_B would have a qualitatively similar effect on D^* and thereby on dispersion.

Finally, panel (d) shows how vote shares develop when R increases, that is when the tie formation process is more directed. Although the average number of social ties per individual does not change, the vote share distribution is less concentrated than in the baseline specification. When the tie formation process is more directed, the likelihood that individuals with differing opinions become peers is lower. Therefore, fewer individuals change opinions, the scope for coordination is lower and the resulting vote share distribution less concentrated.

Testing for peer effects The model provides a framework to test for peer effects in political opinion formation. If the opinions of an individual's peers change, it predicts that there is a chance for the individual to change her own opinion accordingly. Recalling utility (1.1), this chance should be higher for individuals with more peers changing opinions. To apply this intuition to the data, I observe a sample of individuals who change residential locations. Specifically, I merge individual level panel data on party preferences and residential location with regional level results of federal elections. An individual moving to a new region should be more likely to vote for a certain party if the respective party's vote share in that region is higher (Perez-Truglia, 2018). However, party preferences and vote shares in destination regions will also be correlated if individuals take into account the political landscape (or correlated characteristics) of a region when choosing their location of residence. Regional sorting like this implies that the allocation of movers into destination locations is not random, an assumption that has to be made to obtain unbiased estimates of peer effects. Therefore, much of Section 1.5, the empirical part of the paper concerned with testing for peer effects, will focus on establishing and checking the conditions under which changes of residence can be viewed as quasi-random. Additionally, I exploit the model's prediction that individuals with more social ties are exposed to peer effects from more sources. Using the intuition that clubs and associations facilitate face-to-face interactions, I can proxy an individual's number of physical social ties, and thereby their exposure to peer effects in the destination region. In terms of the model, a higher number of clubs and associations corresponds to higher levels of c_F or lower levels of c_B and should increase the effect

of a party's vote share on the individual's probability to vote for it.

Declining number of face-to-face interactions Comparing panels (a) and (c) in Figure 1.2 indicate that two regions inhabited by individuals with different levels of steady state degrees will exhibit (cross-sectional) differences in the concentration of their vote share distributions. Figure 1.3 (a) shows results of simulating the model with the same (baseline) parameters as those used for Figure 1.2 (a) except for a shock reducing c_F from 0.01 to 0.005 from period $t = 151$ onward. Panel (b) repeats the analysis with a stronger peer effect of $\delta = 0.5$. The lower graphs of each panel show that steady state degrees within the same region change from higher to lower levels as response to the negative shock to c_F. At the same time, shown in the upper graphs, voters switch from the majority party to the minority parties, dispersing the vote share distribution. These plots can be viewed as stylized versions of the real vote share and Herfindahl index trends in Germany, shown in Figure 1.1.[18]

When the steady state number of social ties falls as a response to the shock in c_F, individuals are freed from social influence, increasing the relative weight of preference parameters in their utility functions. The expected vote share distribution moves closer to the state without peer effects which is, since preferences are identically distributed, very dispersed.[19]

Based on the same intuition, the model will predict different levels or *changes* in the dispersion of vote shares across or *within* regions. This observation is important because I try to attribute changes in the German vote share distribution to changes in the number of citizen's face-to-face interactions, however, due to data limitations, I will not be able to test this hypothesis with a panel data approach. Instead, I have to rely on a cross-sectional analysis. Using the number of clubs and associations in a region as proxy for face-to-face interactions and the number of social ties between voters in a region, I expect those regions with more clubs and associations to exhibit a higher Herfindahl index of vote shares.

Introduction of broadband internet The introduction of broadband internet has enabled users to interact via online social media and form virtual ties. Especially in the context of partisan affiliation, research has identified that these ties are very likely to form between like-minded individuals. Therefore, I capture the introduction of broadband internet in my model with a positive shock to R. Figure 1.4 (a) shows results of

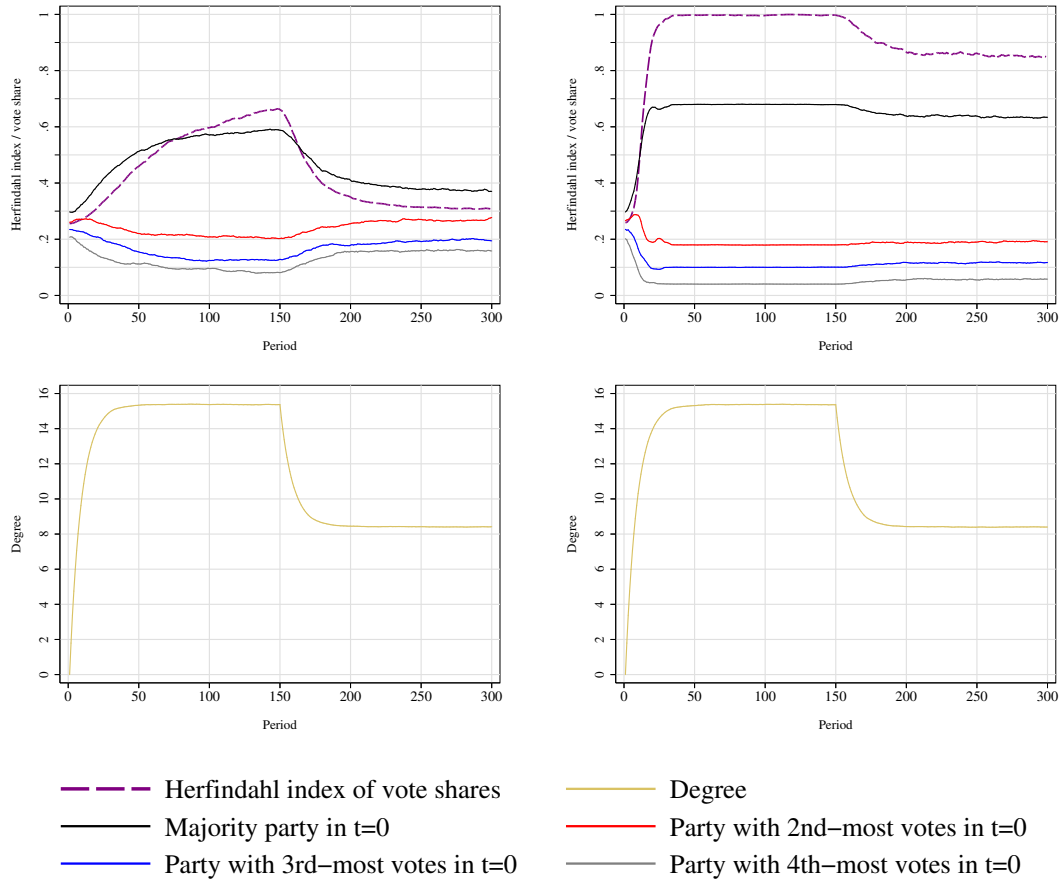
[18] Figure 1.B.4 in the appendix additionally shows exemplary trajectories based on only one simulation and the same parameters. Compare also Footnote 6.

[19] Note that these dynamics depend on the implicit assumption that social influence does not have persistent effects, that is it does not affect preferences. This view is consistent with peer effects driven by Identification (Kelman, 1958).

Figure 1.3: Negative shock to c_F: Simulation results for Herfindahl index, vote shares and degrees

(a) ***Parameters:***
$c_f = 0.01$ for $t \leq 150$ and $c_f = 0.005$ for $t > 150$,
$c_B = 0.05$, $R = 1.5$,
$\delta = 0.2$

(b) ***Parameters:***
$c_f = 0.01$ for $t \leq 150$ and $c_f = 0.005$ for $t > 150$,
$c_B = 0.05$, $R = 1.5$,
$\delta = 0.5$



Notes: This figure shows the average trajectories of vote shares, the Herfindahl index and degrees across 50 simulation loops of the model presented in Section 1.3. Note that the lower panels plot averages across simulations of $d^* = \frac{D^*}{N}$ which is itself an average across individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals, $c_B = 0.05$, $R = 1.5$ and $c_F = 0.01$ for $t \leq 150$ and $c_F = 0.005$ for $t > 150$. $\delta = 0.2$ in panel (a) and $\delta = 0.5$ in panel (b).

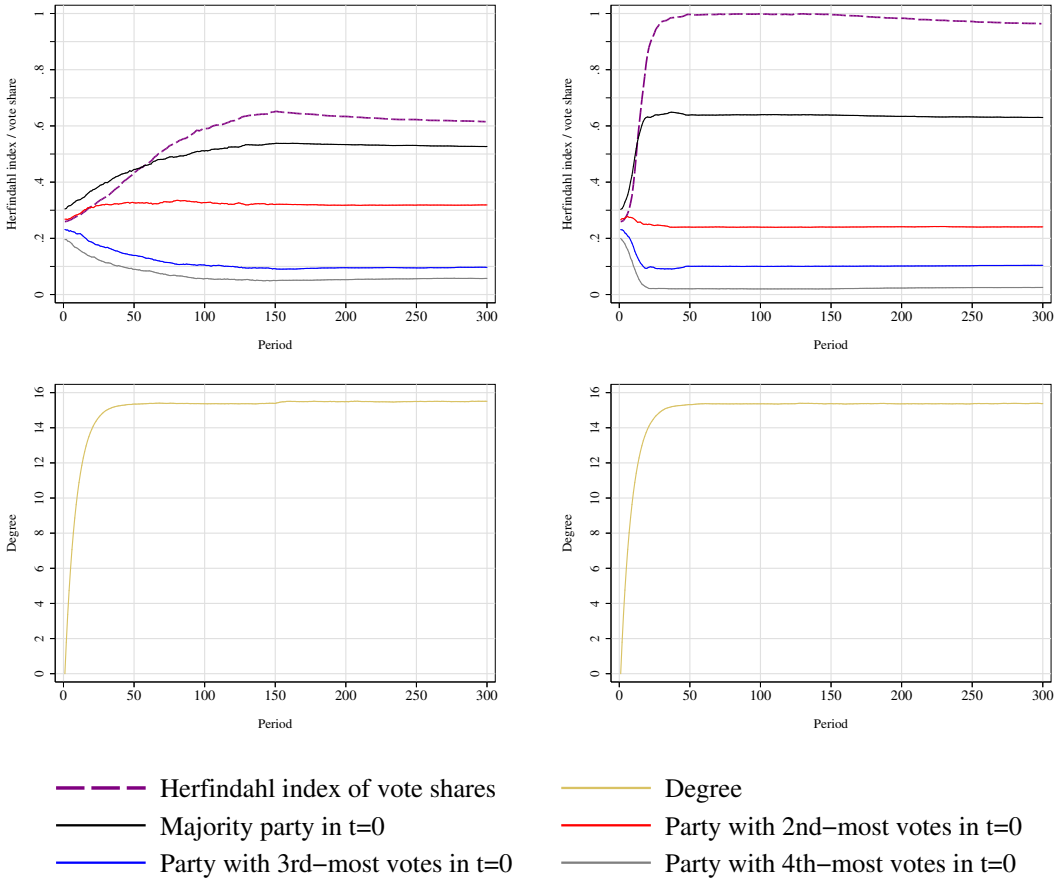
simulations with the baseline parameters used for Figure 1.2 (a) except for a shock increasing R from 1.5 to 5 from period $t = 151$ onward. Panel (b) repeats the analysis with a stronger peer effect of $\delta = 0.5$.[20] Making tie formation more directed leaves

[20] Figure 1.B.5 in the appendix additionally shows exemplary trajectories based on only one simulation and the same parameters.

Figure 1.4: Positive shock to R: Simulation results for Herfindahl index, vote shares and degrees

(a) ***Parameters:***
$c_F = 0.01$, $c_B = 0.05$,
$R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$,
$\delta = 0.2$

(b) ***Parameters:***
$c_F = 0.01$, $c_B = 0.05$,
$R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$,
$\delta = 0.5$



Notes: This figure shows the average trajectories of vote shares, the Herfindahl index and degrees based on 50 simulation loops of the model presented in Section 1.3. Note that the lower panels plot averages across simulations of $d^* = \frac{D^*}{N}$ which is itself an average across individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals, $c_F = 0.01$, $c_B = 0.05$ and $R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$. $\delta = 0.2$ in panel (a) and $\delta = 0.5$ in panel (b).

degrees unaffected, however, since individuals with differing opinions are less likely to meet, there is less scope for coordination and the average Herfindahl index decreases.[21]

[21] One could argue that the possibility of forming virtual ties might enable individuals to have more social contacts. Then, modelling the introduction of broadband internet would involve an additional positive shock to c_F which, as seen before, concentrates the vote share distribution. My empirical results indicate that the net effect of more virtual ties diesperses the vote share distribution.

To test this prediction in the data, I proxy the number of virtual ties individuals can maintain with the broadband penetration rate in their municipality of residence. According to the model, regions in which broadband technology was rolled out faster should have exhibited a faster increase in political fragmentation.

1.3.1. A note on compositional effects

So far, I have explained the observed dispersion of the vote share distribution with exogenous changes in steady state degrees and the direction of the tie formation process within a constant population. Figures 1.3 and 1.4 visualize this intuition. In reality, however, the population constantly changes as new generations are born and the old ones die. Table 1.A.3 in the appendix indicates that it is especially the young who substitute face-to-face interactions with virtual social ties. Therefore, the question arises whether the observed dispersion is instead driven by a compositional effect. Consider an extreme case where both generations are completely separated and therefore could be represented by two different networks in my model. The network of the old generation would exhibit a concentrated vote share distribution (because of many physical social ties) and the network of the young would have a rather diversified political spectrum because the young are more likely to form ties with like-minded others. As the old generation slowly dies, the overall vote share distribution would move towards the vote share distribution of the young, i.e. become more dispersed. Importantly, my model encompasses both intuitions because, as previously mentioned, it works in the cross-section, that is across different networks with different but time-invariant values for c_F, c_B and R and within one network with changing parameter values. The upper panel of Figure 1.B.7 in the appendix plots the development of the federal level Herfindahl index since 1972 split by age groups.[22] In almost every election year, the vote share distribution is more dispersed for younger individuals. The lower panel of the same figure shows a plot of the Herfindahl index split by three different birth cohorts.[23] The Herfindahl index is decreasing for all shown cohorts. So evidence seems to indicate that the overall dispersion of vote shares is due to two reasons. First, the vote share distribution among the younger generations is more dispersed than that of the old. Second, for given birth cohorts, dispersion also increases over time.

[22]The graph starts in 1972 because the definition of age groups in the source data is inconsistent over time.

[23]I recover the cohorts for the lower panel of Figure 1.B.7 from the same data I used for the upper panel of the respective figure which is based on age groups rather than birth years. Unfortunately, the representative survey on voting results by age groups was not conducted in 1994 and 1998. Therefore, I can only recover three cohorts from the data.

1.4. Data

This section provides an overview of the data used throughout this paper. To test for peer effects in political opinion formation, I use individual level and regional level data. To estimate the effects of changes in physical and virtual social ties, I require only regional level data.

All individual data are taken from the Socio-Economic Panel (SOEP) and are available from 1984 to 2014. The SOEP is an annual survey covering a representative share of German households. It is maintained by the German Institute for Economic Research. The dataset follows the same individuals since 1984 and contains a stable set of individual and household characteristics that are surveyed annually or at least periodically. It allows identifying single individuals as well as their municipality of residence. For my analysis, I use data on gender, migration background, partisan affiliation and support, and an individual's location of residence.

Most regional level data are taken from the website of the Federal Statistical Office of Germany and are available only from 2008 onward. I use data on municipality size, business tax multipliers and revenues[24], population (split by gender, age groups, nationalities and education) and unemployment rates.[25] Upon request, I also received municipality level data on population, broken down by age groups and gender for the years 1993 and 2001. For these years, however, data are missing for the federal states of Baden-Württemberg and Saarland.

Voting data are taken from the website of the *Bundeswahlleiter*, the person (usually the head of the Federal Statistical Office of Germany) mandated to oversee that federal elections take place according to the legislation. For the Herfindahl index to be consistent over time, I base its calculation on those three parties eligible for federal elections since 1980: The CDU, the SPD and the FDP. I group vote shares of all other parties into one remainder category. For regressions using the Herfindahl index as dependent variable, I provide robustness checks in which the Herfindahl index is additionally based on the vote share of the Green party.

To proxy the number of physical social ties between individuals in a region, I obtained data on the number of clubs and associations on the municipality level from the German registry of clubs and associations (*Vereinsregister*). It is an online portal list-

[24]Each municipality in Germany can multiply the fixed business tax rate with a multiplier (*Hebesatz*) to adjust effective tax rates.

[25]The Federal Statistical Office does not provide municipality level data on unemployment rates but only on the number of unemployed (and the Federal Employment Agency publishes unemployment rates only on the county level). I approximate the unemployment rate by dividing the number of unemployed through the number of individuals aged 18 to 65. The official definition would require using the labor force as divisor.

ing all registered German clubs and associations alongside with some basic information on them (e.g. location). Note that this list is not intended for research purposes but is kept because of legal reasons. Therefore, it only represents a current cross-section of the landscape of associations in Germany. After obtaining the registry's permission, I was able to collect the relevant data from their homepage in May 2019.[26] Although frequently used in the literature (compare Sections 1.1 and 1.2), there are some issues to be kept in mind concerning this variable. First, since my panel data on party support is only available until 2014, an individual's face-to-face interactions and number of physical ties might not be adequately captured in my individual level estimations. I try to alleviate this problem by only considering an individual's last residential change and by presenting results for subsamples of late movers. Implicitly, I have to assume that at least relative differences between regions in terms of physical social ties remained relatively constant over time. A second problem is that a higher density of clubs and associations within a region might proxy for a very fragmented society. If there is a club for each political opinion (i.e. political networks), peer effects will most likely operate within these subnetworks and not be measurable across them, i.e. on the regional level. In terms of the model this means that clubs and associations may not increase the probability that individuals with different opinions meet. My findings, however, are in line with my model's prediction, indicating that clubs and associations do indeed increase the probability that individuals with different political opinions meet. Using clubs and associations as measure for individual degrees also has advantages compared to otherwise appropriate variables. Measuring social ties with, for example, an individual's number of friends in a region could create endogeneity problems. An individual might be more likely to make new friends in a region where more like-minded others live. Since clubs and associations are a regional level variable, this type of reversed causality should pose no threat to identification.

Data on the location of district courts, the instrumental variable for clubs and associations, is taken from the *Anwaltsverzeichnis für Deutschland*, an online registry to locate courts and law firms which I accessed in 2021. A detailed discussion of the relevance and exogeneity of the instrument is deferred to Section 1.6.

Information on broadband penetration for the years 2005 to 2008 are taken from Falck, Gold, and Heblich (2014). Residents of larger German cities were able to access the internet via a broadband connection for the first time in 1999. In mid 2002, the technology started to become available for a broader public. Therefore, broadband penetration was zero before 1999 and still close to zero before 2003. To instrument for the regional broadband penetration, I use information on the distance of a municipality

[26] I scraped the respective data using Python libraries *BeautifulSoup* and *Mechanize*.

to its so-called *Main Distribution Frame* (MDF), a physical structure whose location was relevant for households' bandwidth in the early stages of broadband technology. Location data are also taken from Falck, Gold, and Heblich (2014) and a more detailed discussion of the variable can be found in Section 1.7.

Finally, for Tables 1.A.2 (further dispersing opinions in Germany) and 1.A.3 (development of face-to-face and virtual interactions) in the appendix, I use results from several surveys based on representative samples of the German population. Trends on dispersing opinions are taken from surveys conducted by the Allensbach Institute (*Institut für Demoskopie Allensbach*), a private polling institute. Data on face-to-face and virtual interactions stem from two different sources. First, the German General Social Survey (*Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS*), conducted by the Leibniz Institute for the Social Sciences, a major research institute funded by the German government. Second, a survey on how the German population allocates its time across different activities, conducted by the Federal Statsistical Office of Germany.

Table 1.C.1 in the appendix provides an additional overview of the data, including measurement scales, units and the respective sources. Summary statistics for all variables relevant for estimations are provided in the respective sections.

1.5. Testing for Peer Effects

1.5.1. Empirical strategy

This section is intended to test the basic assumption of my model that voters are prone to peer effects when deciding which party to vote for. Intuitively, if a considerable fraction of an individual's reference group changes their voting decisions in the same way, the individual herself is likely to change her own opinion accordingly. This likelihood increases with the number of peers changing their voting decisions. To test this hypothesis in the data, I observe a sample of movers and their stated partisan affiliation. The variation in the peers' voting decisions stems from the change in vote shares resulting from the change in residential locations. Variation in the number of peers the individual is likely exposed to is based on the variation in the number of clubs and associations between different residential locations. In this context, an individual who would be allocated randomly to a new location should be more likely to vote for a party the higher is the vote share of that party in the new location. This effect should be stronger for individuals moving to regions with many clubs and associations because they foster the creation of physical social ties.

My individual level data allows me to identify movers and track their locations.

Most importantly, for each survey year, it provides information on the political party supported and the degree of that support. The annual survey used for the data collection process allows individuals to choose exactly one out of a comprehensive list of German political parties as their favourite. In another question, the support for that party is elicited on a scale from 1 to 5. I can match these data with regional characteristics, in particular the respective vote shares on the municipality level and then implement the above mentioned approach. I begin my analysis by collapsing the person-year level data to the person-period level. Each period is defined as the time between two residential moves (across municipality borders) or the time between the first entry of an individual into the panel and her first residential move or the time between her last residential move and the last time she was observed in the panel. Say a person lived in municipality A from 1984 to 1990 (possibly changing residence within that municipality) and in municipality B from 1991 to 2000. What would have been 17 observations for this individual in the original data are now 2 observations in the new data. Unless stated otherwise, all variables mentioned from now on will be averages across the years comprising a period. For example, the vote share of any party in municipality A for the above exemplary individual would be its average vote share across the years 1984 to 1990. As in Perez-Truglia (2018), I only consider the last move for those individuals who moved several times. For clarity, denote the pre-move period by t_{i0} and the post move period as t_{i1}. The time dimension of the panel data will consist only of those two periods. The periods are indexed by i because the length and time of the periods vary for each person. For each individual, I observe the party she most often supported in period t_{i0}, P_{i0}, and the extent of that support S_{i0}. Consider again the exemplary individual i who moved from municipality A to B. Say, from 1984 until 1988 she supported the SPD and in 1989 and 1990 the Green party. When answering the question eliciting party support, she quantified the extent of her support with numbers 1, 2, 3, 3 and 4 in the years she supported the SPD, respectively and with numbers 2 and 3 for the last two years in which she supported the Green party. Since the SPD is the party she supported most often before her move, for this individual, I set P_{i0} ="SPD" and $S_{i0} = \frac{1+2+3+3+4}{7} = 2.5$.[27]

Let S_{i1} be individual $i's$ support of P_{i0} in period t_{i1}, i.e. S_{i1} is the degree of support in period t_{i1} for the party an individual used to support most often in period t_{i0}. To quantify S_{i1} for the exemplary individual i, I check how often she supported the SPD in the years from 1991 to 2000. If she did not support the SPD in any of those years, my original data set would report missing values for her party support in each of those years because the extent of party support is only elicited for one party, the supported party. In

[27] Implicitly, I am assuming that i's support for the SPD is zero, instead of missing, for the years 1989 and 1990.

this case I decide to set $S_{i1} = 0$. While S_{i0} can take any value between 1 and 5, S_{i1} can take any value between 0 and 5. If the individual does support the SPD in any of the years she lives in municipality B, I add up the party support elicited in those years and divide it by the number of years, I observe her living in B, which would be 10 years in this example.

Finally, denote the vote share of P_{i0} in individual i's destination municipality (that is in t_{i1}) as VS_{i1} and the number of clubs and associations in that municipality as CA_{i1}. For the exemplary individual, VS_{i1} will be the average vote share of the SPD in municipality B, taken across the years ranging from 1991 to 2000. Based on VS_{i1}, I construct a dummy variable, MJ_{i1}, which indicates whether or not P_{i0} is the majority party in t_{i1}, i.e. whether the party an individual supported most often before her move has, on average, the highest vote share among all parties in the individual's post-move municipality. It is important to note why I do not use the vote share of P_{i0} in the destination region, VS_{i1}, but the dummy variable MJ_{i1}. More clubs and associations imply more social ties which, according to my model, increase the magnitude of peer effects. However, they increase the magnitude of peer effects for all parties. Therefore, it is not clear whether the effect of the vote share of P_{i0} in the destination region, VS_{i1}, should be stronger for an individual with more social ties in the destination region. For example, if tie formation was close to random, and P_{i0} has a very low vote share in the destination region, it is likely that newly arrived individuals form ties with peers who hold different opinions than P_{i0}. In that case, the effect of VS_{i1} on the probability to vote for P_{i0} would be weaker in regions with many clubs and associations. It will be unambiguously stronger only if P_{i0} is indeed the majority party in the destination region. Then, even under random tie formation, newly formed ties are most likely to have formed with other individuals who also vote for P_{i0}.

With this intuition in mind, I estimate the following baseline models (based on Perez-Truglia, 2018):

$$S_{i1} = \alpha_0 + \alpha_1 MJ_{i1} + \alpha_2 MJ_{i1} * CA_{i1} + \alpha_3 CA_{i1} + \boldsymbol{X_i'a} + u_i \tag{1.8}$$

$$S_{i0} = \beta_0 + \beta_1 MJ_{i1} + \beta_2 MJ_{i1} * CA_{i1} + \beta_3 CA_{i1} + \boldsymbol{X_i'b} + v_i \tag{1.9}$$

Equation (1.8) represents the regression of an individual's post-move support of the party she used to support most often before changing residence on the dummy indicating whether that party won the majority of votes in the destination region, interacted with the number of clubs and associations in that region. The interaction between MJ_{i1} and the number of clubs and associations, CA_{i1}, is intended to capture the higher effects of a party's vote share on individual party support when an individual has more social

ties. Equation (1.9) is specified to check whether regional sorting might bias the results of regression (1.8). It represents a regression of the pre-move party support of P_{i0} on that party's majority status in the destination region, MJ_{i1}. If, conditional on controls, individuals chose residential locations based on their political landscape or related attributes, one would expect the majority status of P_{i0} in the destination municipality to be positively correlated with an individual's support for it in the origin region, S_{i0}. $\boldsymbol{X_i}$ is a vector of controls, including, amongst others, the *interaction* of three fixed effects: a dummy indicating which party, P_{i0}, the individual supported before her move. Another dummy to capture different levels of the vote share of P_{i0} in her origin municipality (in steps of 0.01). Finally, a dummy variable with ten values grouping individuals according to their levels of support for P_{i0} in the origin municipality, S_{i0}. These interactions partition the sample of the first and second regression into groups of individuals supporting the same pre-move party and arriving from municipalities with similar (political) characteristics. Including these dummies makes sure that only similar individuals are compared such that the choice of destination locations among them can be assumed to be conditionally exogenous to regional political characteristics. For regression (1.8), $\boldsymbol{X_i}$ further includes the origin region party support S_{i0}, measured on its original scale. Additionally included in both specifications are the following destination region characteristics: the size of the municipality (in square kilometres), a cubic of its population, the share of the population older than 65, the share of the population younger than 18, the share of individuals who are non-German, the share of individuals without any vocational education, the Herfindahl index of foreign nationalities, the local business tax multiplier and revenue and the unemployment rate. Data on nationalities and education are only available at the county level. I also include dummies controlling for gender and a possible migration background.[28] Standard errors are clustered at the municipality level. Table 1.1 shows the respective summary statistics.

If peer effects are important for political opinion formation, α_1 should be positive. If peer effects are stronger when individuals have more social ties, as predicted by the model, α_2 should also be positive. If, conditional on the included control variables, sorting is not a problem β_1 and β_2 should be insignificant.

1.5.2. Results

Columns (1) and (2) of Table 1.2 show the coefficients of regression Equations (1.8) and (1.9), respectively. Column (3) repeats the estimation of Equation (1.8) but with a different dependent variable: a dummy indicating whether an individual elected the same party in her destination region as in her origin region. While it is not possible

[28] I choose control variables based on the specifications of Perez-Truglia (2018).

Table 1.1: Summary statistics, testing for peer effects

Variable	Post-move sample		Pre-move sample	
	Mean	Std.Dev.	Mean	Std.Dev.
Supported party				
CDU/CSU (Conservatives)	0.39	0.49	0.36	0.48
SPD (Social Democrats)	0.37	0.48	0.37	0.48
FDP (Liberals)	0.03	0.18	0.04	0.20
B90 Die Grünen (Green Party)	0.12	0.33	0.13	0.34
Die Linke (The Left)	0.09	0.28	0.09	0.29
Degree of party support (*t=1*)	2.96	1.15		
Degree of party support (*t=0*)	3.29	0.69	3.21	0.72
Elected same party	0.89	0.31		
Supported party won majority (*t=1*)	0.40	0.49	0.40	0.49
Vote share (*t=1*)	0.29	0.12	0.28	0.13
Vote share (*t=0*)	0.31	0.13	0.30	0.13
Clubs&Associations (*in 1000*)	0.89	1.97	0.85	1.92
Area (*in km^2*)	98.42	91.62	97.12	91.10
Population (*in 1000*)	112.46	220.28	107.76	215.52
Share of population older than 65	0.21	0.04	0.22	0.04
Share of population younger than 18	0.16	0.03	0.16	0.03
Share of foreigners	0.05	0.05	0.05	0.05
Share of non-educated	0.11	0.04	0.11	0.04
Herfindahl index of foreigners	0.08	0.05	0.08	0.05
Business tax multiplier	3.83	0.53	3.82	0.53
Business tax revenue (*in k€*)	82.45	249.17	79.78	247.90
Unemployment rate	0.07	0.03	0.07	0.03
Individual is male	0.49	0.50	0.52	0.50
Individual was born abroad				
No	0.86	0.35	0.84	0.36
Yes	0.06	0.24	0.07	0.26
No - but parents	0.08	0.27	0.09	0.28
Observations	2437		3961	

Notes: This table shows summary statistics for all variables used in Equations (1.8) and (1.9). If not mentioned explicitly, all variables refer to the destination location ($t = 1$). The sample sizes differ because for some individuals post-move data are missing.

to test for regional sorting using this dependent variable, its interpretation is easy and the regression serves as further evidence for the existence of peer effects. Note that the clubs and associations variable is demeaned prior to estimation, so that the resulting coefficients on the main effects of the MJ_{i1} regressors represent the corresponding marginal effect at the average municipality in terms of the number of local clubs and associations. The last line of Table 1.2 shows into how many groups the regression sample is partitioned by the interacted fixed effects.

Results are in line with the predictions of the model. The coefficients of specifications (1) and (3) show that higher vote shares for a party, captured by the MJ_{i1} dummy variable, increase an individual's party support in the destination region and the proba-

Table 1.2: Estimation results, testing for peer effects

Dependent variable:	Party support		Elected same party
Pre- / Post-move:	Post	Pre	Post
	(1)	(2)	(3)
Regressors:			
Majority=1	0.636***	0.010	0.222***
	(0.069)	(0.013)	(0.020)
Majority=1×Clubs&Assoc.	0.048***	-0.007	0.014**
	(0.018)	(0.005)	(0.006)
Mean of dependent variable	2.96	3.21	0.89
Obs.	2437	3961	2437
R^2	0.427	0.884	0.338
# of groups	420	517	420

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression Equations (1.8) and (1.9) in columns (1) and (2), respectively. Column (3) shows results of a regression similar to that of column (1) but uses a dummy indicating whether an individual voted for the same party before and after moving as dependent variable. Control variables include the destination municipality's size, its population (as a cubic), its population shares of individuals older than 65, younger than 18, of foreigners, of non-educated, its Herfindahl index of foreigner shares, its business tax multiplier and revenue, its unemployment rate, individual gender and migration background. For the post-move regressions, controls additionally include individual party support in the origin region. Each regression also includes the following, interacted fixed effects: An indicator for the supported party before moving, an indicator for different levels of party support and an indicator for different levels of the vote share of the supported party in the origin region. Number of groups indicates into how many groups the sample is partitioned by the interacted fixed effects. Standard errors are clustered at the municipality level.

bility to vote for the same party as in the origin region. When P_{i0} is the majority party in the destination region, individual party support is 0.6 points higher than in cases where it is one of the minority parties. This is equivalent to roughly 20% of the average, and 55% of the standard deviation of the party support variable in the destination region. For each additional 100 clubs and associations above the average number of 890 per municipality, the effect increases by 0.005. Put differently, a standard deviation increase in the number of clubs and associations above the average corresponds to an increase in effect size by about 8% of a standard deviation. An individual's probability to vote for the same party as in the origin location is 22 percentage points higher if P_{i0} is the majority party in the destination region. This corresponds to 25% of the average individual probability to vote for the same party twice, or 71% of a standard deviation. Again, the effect increases for municipalities with an above average number of clubs and associations. In regions with an additional 100 clubs and associations above the average, the effect increases by 0.001 which is equivalent to a standard deviation increase causing a 9% of a

standard deviation increase in effect size. The results of column (2) indicate that, conditional on the control variables included in the specification, regional sorting is unlikely to explain the significant coefficients of columns (1) and (3). If individuals were to choose destinations based on their political landscape or correlated characteristics, an individual's party support of P_{i0} in the origin region should be significantly correlated with the vote share or the majority status of that party in the destination region. This is clearly not the case.

The regression results show how peer effects will lead to a concentration of the vote share distribution. Individuals are drawn away from supporting minority parties and pushed towards voting for the majority party in the destination region. Clearly, shifting votes between parties in this manner increases the Herfindahl index, i.e. the concentration of the vote share distribution. The positive coefficient on the interaction term highlights that, as predicted by the model, more social ties between voters lead to even more concentration.

Recall that I observe the clubs and associations variable only in 2019 so that face-to-face interactions of early movers might not be adequately captured. To address this problem, I repeat the estimations shown in Table 1.2 with different subsamples including only late movers. The results for including only individuals who moved after 2000, 2005 and 2010 are shown in Tables 1.D.1 to 1.D.3 of Appendix D, respectively. Results are robust throughout the specifications. Only the coefficient on the interaction term of the model shown in column (1) loses some significance when restricting the sample to residential changes between 2010 and 2014. However, this is likely to due the small sample size of 316.

1.6. Explaining Dispersion: Physical Social Ties

1.6.1. Empirical strategy

In this section, I test the first key predictions of my model. The decrease in the concentration of the vote share distribution can be explained with a contemporaneous decline in the number of physical social ties between individuals in a region. Since I observe the number of registered clubs and associations only in 2019, I cannot assess the effect of *changes* in the number of physical social ties. However, my model also predicts that, everything else equal, regions with higher steady state degrees (i.e. more clubs and associations) should exhibit a higher Herfindahl index. Since cross-sectional regressions are prone to endogeneity, the baseline results will be of correlational nature. In the second part of this section, I will use an instrumental variable to test whether there also

exists a causal relationship as predicted by the model.[29]

Baseline specification Equation (1.10) describes the baseline empirical specification to test the model's prediction. I regress the Herfindahl index of vote shares in 2017, $H_{m,2017}$, in municipality m on a constant (β_0), a county fixed effect (β_c), the number of clubs and associations (in thousands) in municipality m in 2019 ($CA_{m,2019}$) and a vector of control variables $\boldsymbol{X}$. Standard errors are clustered at the municipality level.

$$H_{m,2017} = \beta_0 + \beta_c + \beta_1 CA_{m,2019} + \boldsymbol{X}'_{\boldsymbol{m,2017}}\boldsymbol{b} + \epsilon_{m,p} \tag{1.10}$$

$\boldsymbol{X}_{\boldsymbol{m,2017}}$ contains the following variables: the size of a municipality (in square kilometres) and a cubic of its population since both are clearly related with the regressor and, in general, with the outcome variable. Densely populated regions tend to exhibit higher levels of political dispersion. The female population share, the share of the population older than 65, the share of the population younger than 18 and the Herfindahl index of age group shares since any type of local heterogeneity within the population affects election results and potentially the population's willingness or ability to found clubs and associations. The local business tax multiplier and revenue and the unemployment rate because regional economic conditions might affect social participation and election outcomes. Any other type of regional heterogeneity I address with the county fixed effect because I do not have access to other variables capturing further dimensions of heterogeneity at the municipality level. Finally, since the number of physical ties within a region might be (negatively) correlated with the number of virtual ties, I also include a municipality's distance to its MDF, the instrument for broadband penetration.

Instrumental variable specification Although I can control for some potential sources of endogeneity in the above specification, there remain threats to identification. As already mentioned, many types of local heterogeneity might be omitted variables. Also, the causal interpretation of the baseline coefficients could be affected by a reversed causality problem. In a more politically homogeneous society, individuals might find it easier to found and register clubs and associations, leading to a positive correlation between the concentration of the vote share distribution and the numbers of clubs and associations in a region. To tackle these concerns, I implement an instrumental variable approach which is intuitively very similar to the *near-college* approach proposed by Card (1995). Any club or association in Germany has to be registered at the so-called *Amtsgericht*, a district court with the lowest authority within the German judiciary sys-

[29] All of the specifications in Section 1.6 yield similar results with respect to significance and signs when using the number of clubs and associations per capita as main regressor.

tem.[30] As of 2021, out of the roughly 11,000 German municipalities, 688 had their own district court. The modern, uniform (across then independent regions) German judiciary system, and the district courts as one of its part, were established in 1879 through a series of laws called the *Reichsjustizgesetze*. Since then, however, some municipalities have gained and some have lost district courts. Still, most district courts were established and their jurisdictions chosen long before 2017.[31] Intuitively, the relevance of the instrument results from the fact that it is more costly to register an association in a municipality that does not have its own district court. Exogeneity requires that the location and jurisdiction of the district courts is not related to political dispersion in any other way than through their effects on the number of clubs and associations within a region. Since location and jurisdiction were decided well before 2017, contemporaneous shocks and reversed causality pose no threat to the exogeneity of the instrument. However, there are two other concerns regarding identification. First, location characteristics which influenced the probability of a district court being established in a municipality might still have an effect on political dispersion today. It is likely that district courts were established in more agglomerated, economically stronger and populated cities. Assuming that these characteristics are relatively persistent over time renders the instrument endogenous because agglomerated cities tend to exhibit a more dispersed political spectrum. Second, the presence of a district court (and its jurisdiction) could affect political dispersion through other channels than its effects on the number of clubs and associations. Presumably, the establishment of a district court changed the economic performance of a municipality which has consequences for the political landscape. Using the control variables of the baseline specification, I try to ensure the conditional exogeneity of my instrument.

The respective two-stages-least-squares (2SLS) regressions I am implementing are as follows:

$$CA_{m,2019} = \beta_0 + \beta_c + \beta_1 AG_{m,2021} + \boldsymbol{X}'_{\boldsymbol{m,2017}}\boldsymbol{b} + v_{m,p} \tag{1.11}$$

$$H_{m,2017} = \alpha_0 + \alpha_c + \alpha_1 \widehat{CA_{m,2019}} + \boldsymbol{X}'_{\boldsymbol{m,2017}}\boldsymbol{a} + u_{m,p} \tag{1.12}$$

where Equation (1.11) is the first stage, $AG_{m,2021}$ is an indicator equalling unity if a municipality had its own district court in 2021 and zero else and $\widehat{CA_{m,2019}}$ are the predicted values of the clubs and associations variable. Standard errors are clustered at the municipality level. Table 1.3 contains summary statistics for all variables used in the

[30] Typically, only cases with expected imprisonment under four years are tried in those courts.

[31] The youngest district court was established in 1929 in *Singen*. Following a legislative reform in 2013, ten district courts in the federal state of *Mecklenburg-Vorpommern* were shut down, the last one in *Ribnitz-Damgarten* in 2017. An additional district court was shut down in 2016 in *Gelsenkirchen-Buer*.

Table 1.3: Summary Statistics, physical ties

Variable	Mean	Std.Dev.	Min	Max
2017				
Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.31	0.03	0.25	0.60
Based on CDU, SPD, FDP and Green party	0.28	0.04	0.20	0.60
Clubs&Associations (*in 1000*)	0.03	0.07	0.00	2.20
Municipality has own district court	0.05	0.23	0.00	1.00
Area (*in* km^2)	28.38	30.65	0.39	303.10
Population (*in 1000*)	5.66	10.45	0.07	180.97
Female population share	0.50	0.02	0.26	0.60
Share of population older than 65	0.21	0.04	0.10	0.54
Share of population younger than 18	0.17	0.02	0.05	0.28
Herfindahl index of age groups	0.46	0.02	0.37	0.64
Business tax multiplier	3.61	0.34	2.30	5.50
Business tax revenue (*in k€*)	3.25	10.35	-11.60	301.64
Unemployment rate	0.03	0.01	0.00	0.16
Distance to MDF (*in km*)	2.88	1.75	0.03	14.83
Observations		7791		

Notes: This table shows summary statistics for all variables used in Equations (1.10), (1.11) and (1.12).

baseline as well as IV specifications.[32]

1.6.2. Results

Table 1.4 shows the coefficients of interest of specifications (1.10), (1.11) and (1.12) in columns (1), (4) and (2), respectively. Additionally, column (3) presents results of the reduced form regression of the Herfindahl index in 2017 on the instrumental variable. Generally, municipalities with more clubs and associations have more inhabitants and are more agglomerated. It is these regions which exhibit dispersed vote share distributions. Therefore, the estimates in the baseline specification likely suffer from a downward bias. Still, as predicted by the model, the respective coefficient of the clubs and associations variable is significant and positive. Municipalities with more clubs and associations exhibit a more concentrated vote share distribution. For each 100 clubs and associations, the 2017 Herfindahl index increases by 0.003 which is equal to 1% of the average Herfindahl index across municipalities or 10% of a standard deviation. Analogously, a standard deviation increase in the number of clubs and associations is related to an 8% of a standard deviation increase in the 2017 Herfindahl index. The IV

[32]The differences in summary statistics between regional level variables of this sample and the ones used in the previous section are due to the fact that, in this section, I do not merge regional and individual level variables. Individual level data from the SOEP is based on individuals which tend to live in bigger cities so that a lot of smaller regions are missing from that analysis. Here, these municipalities are not missing.

Table 1.4: Estimation results, physical ties

Dependent variable:	Herfindahl index 2017			Clubs&Assoc.
Specification:	Baseline	IV	Reduced	1st stage
	(1)	(2)	(3)	(4)
Regressors:				
Clubs&Assoc.	0.033***	0.071***		
	(0.006)	(0.020)		
Municipality has district court			0.002***	0.035***
			(0.001)	(0.003)
Population	-0.001***	-0.001***	-0.001***	0.005***
	(0.000)	(0.000)	(0.000)	(0.001)
Unemployment rate	-0.152***	-0.159***	-0.151***	0.101***
	(0.028)	(0.028)	(0.028)	(0.038)
Mean of dependent variable	0.31	0.31	0.31	0.03
Obs.	7791	7791	7791	7791
R^2	0.712		0.711	
Kleibergen-Paap rk Wald F statistic				129
Anderson-Rubin Wald F statistic				12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regressions Equations (1.10), (1.11) and (1.12) in columns (1), (4) and (2), respectively. Column (3) shows the reduced form regression of the IV specification in column (2). Control variables include a municipality's size, its population (as a cubic), the female population share, the share of individuals older than 65, younger than 18, its Herfindahl index of age group shares, its business tax multiplier and revenue, its unemployment rate, the distance to its MDF and a county fixed effect. Standard errors are clustered at the municipality level.

coefficient is more than twice as large confirming that, indeed, the OLS estimate of the baseline specification is downward biased. With an F statistic of 129, the instrument is sufficiently strong. Note also that the biggest threat to the exogeneity of the IV coefficient comes from the fact that district courts might be, even conditional on control variables, located in more agglomerated cities. Therefore, if the IV estimate should be biased, it is likely smaller than the real causal coefficient.

These results confirm the predictions of the model that regions inhabited by individuals with more physical social ties exhibit a more concentrated vote share distribution. Together with the outcome of the previous section, these findings provide evidence for one of the main predictions of my model and therefore for the intuition that the observed decline of face-to-face contacts can explain at least parts of the decline in the concentration of the vote share distribution.

1.6.3. Alternative specification of Herfindahl index

As a robustness check, Table 1.E.1 in the appendix repeats the analyses shown in Table 1.4 with the Herfindahl index being based not only on the vote shares of the CDU, SPD and FDP but also the Green party. In this specification, the OLS coefficient drops significantly from 0.033 to 0.005 and looses its significance. Given that the IV coefficient increases slightly from 0.071 to 0.091 and stays significant at the 1% level, however, it seems likely that the reduction of the OLS estimate is due to the downward bias explained in the previous section. Consistent with this view is the fact that the Green party in Germany is especially strong in urban areas which are those with a relatively high number of clubs and associations. Including the Green party's vote share in the calculation of the Herfindahl index will therefore yield a relatively low value in those regions with many clubs and associations, thereby weakening the positive correlation in the OLS regression.

1.7. Explaining Dispersion: Virtual Social Ties

1.7.1. Empirical strategy

In this section, I test the second main prediction of my model, whether the increase in the dispersion of the vote share distribution can be explained by an increase in virtual interactions between voters. Using regional broadband penetration rates to proxy inhabitants' possibilities to form virtual ties, I expect that regions in which broadband was rolled out faster exhibited a higher increase in the dispersion of their vote share distribution.

Baseline specification Internet providers started to roll out broadband technology in agglomerated areas which are systematically different from the rest of Germany. As previously mentioned, these regions tend to exhibit more dispersed vote shares, so that the estimates of a cross-sectional regression of the Herfindahl index on broadband penetration would likely be downward biased. To account for this fixed effect, I employ a first-difference specification where I regress differences in the Herfindahl index within a region on changes in broadband penetration and control variables. I choose to calculate differences in the Herfindahl index of vote shares based on federal election years 2002 and 2009 although the last federal election before the official introduction of broadband technology (in 1999) took place in 1998. Since I observe control variables only for the years 1993, 2001 and from 2008 onward, this choice seems more adequate than using the Herfindahl index of 1998 as base year. Also, it is safe to assume broadband

penetration to be negligible for the years 2000 to 2002 because broadband connections were still a niche technology then. Finally, I use 2008 values for the broadband penetration in 2009 (which I do not observe).[33] Equation (1.13) shows the respective model specification:

$$\Delta H_{m,'09-'02} = \beta_0 + \beta_c + \beta_1 \Delta BP_{m,'08-'01} + \boldsymbol{\Delta X'_{m,'09-'01}} \boldsymbol{b} + \epsilon_m \tag{1.13}$$

Δ is the first-difference operator, $\Delta BP_{m,'08-'01}$ is the difference between broadband penetration in 2008 and 2001 (which equals the 2008 value because broadband penetration in 2001 was negligible) and $\boldsymbol{\Delta X}$ is a vector of control variables. It includes differences of the following variables: a cubic of a municipality's population, the female population share, the share of the population older than 65, the share of the population younger than 18, the Herfindahl index of age groups and the unemployment rate. These controls are intended to capture the socioeconomic status of a municipality. $\boldsymbol{\Delta X}$ also includes the Herfindahl index (level) of 2002 to capture potential effects on changes in political fragmentation which stem from differences in initial levels. The county fixed effect β_c allows for differential linear trends across municipalities belonging to different counties. To rule out that a potential (negative) correlation between the number of physical and virtual ties in a region affects my estimates, $\boldsymbol{\Delta X}$ also includes the dummy indicating whether a municipality has its own district court. Standard errors are clustered at the municipality level.

Instrumental variable specification Despite the first-difference specification, endogeneity is still a concern. For example, agglomerated regions experienced significantly smaller decreases in their Herfindahl indices until 2009. Since it is these regions in which broadband was rolled out the fastest, my first-difference coefficients are likely upward biased.[34]

Therefore, I use an instrumental variable constructed by Falck, Gold, and Heblich (2014) to proxy for changes in broadband penetration. To understand the relevance of the instrument, consider the following paragraph which is based on their paper: when broadband was introduced in Germany, it relied on the existing telephone infrastructure. A bottleneck for the available bandwidth at the household level was the distance of a household to its *Main Distribution Frame* (MDF), a unit (located in a dedicated build-

[33] Instead, I could have calculated differences between 2001 and 2008 values. However the 2008 Herfindahl index would have been based on federal election results from 2005.

[34] Note how the potential bias is of the opposite sign as expected in a cross-sectional setting: while agglomerated regions tend to exhibit more dispersed vote share distributions (downward bias), they exhibit lower (negative) *changes* in dispersion between 2002 and 2009 (upward bias).

ing) interlinking the national telephone network with local street cabinets. The greater the distance, the lower the bandwidth available to a household. To construct their instrumental variable, Falck, Gold, and Heblich (2014) collected data on the distance of all German municipalities (where the location of a municipality is its geographic center) to their MDFs. This variable is potentially exogenous because for telephone service quality the distance of an MDF to its municipality did not matter so that other considerations (in particular available lots) had a bigger impact on location choice. According to the authors, the biggest concern regarding the exclusion restriction is that MDFs were more likely to be located closer to agglomerated areas which exhibit different voting behaviours. The variables of the baseline specification are chosen to control for this possible source of endogeneity. For further explanations on the instrument see Falck, Gold, and Heblich (2014) or Campante, Durante, and Sobbrio (2018) who use a similar approach.

The respective 2SLS specifications are as follows:

$$\Delta BP_{m,'08-'01} = \beta_0 + \beta_c + \beta_1 D_m + \boldsymbol{\Delta X'_{m,'09-'01}} \boldsymbol{b} + v_m \tag{1.14}$$

$$\Delta H_{m,'09-'02} = \alpha_0 + \alpha_c + \alpha_1 \Delta \widehat{BP_{m,'08-'01}} + \boldsymbol{\Delta X'_{m,'09-'01}} \boldsymbol{a} + u_m \tag{1.15}$$

where Equation (1.14) is the first stage. D_m is the distance (in km) of a municipality to its MDF and $\Delta \widehat{BP_{m,'08-'01}}$ are the values of broadband penetration predicted in the first stage. Table 1.5 displays summary statistics of all (differenced) variables used in these specifications. Table 1.F.1 in the appendix contains summary statistics for 2002 (2001) and 2009 (2008) values of the respective variables.

1.7.2. Results

Table 1.6 shows coefficients of regressions (1.13), (1.14) and (1.15) in clumns (1), (4) and (2), respectively. Additionally, column (3) displays coefficients of the reduced form of the IV specifications (1.14) and (1.15). The results are in line with the predictions of the model. A greater increase in broadband penetration is related to a more negative difference of the Herfindahl index between 2009 and 2002, implying that regions in which broadband technology was rolled out faster experienced a greater decrease in the concentration of their vote share distributions. However, effects are rather small. The baseline coefficient in column (1) indicates that a ten percentage point increase in broadband penetration increases the negative difference in the Herfindahl index by 0.0004, which corresponds to about 0.4% of the average decline observed across the

Table 1.5: Summary statistics, virtual ties

Variable	Mean	Std.Dev.	Min.	Max.
Δ2009-2002				
Δ Herfindahl index of vote shares				
Based on CDU, SPD and FDP	-0.11	0.06	-0.38	0.08
Based on CDU, SPD, FDP and Green party	-0.13	0.06	-0.41	0.06
Share of municipalities with decreasing Herfindahl index	0.99	0.07	0.00	1.00
Δ Share of households connected to broadband network	0.90	0.17	0.00	1.00
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Δ Population (*in 1000*)	-0.03	0.41	-5.54	4.57
Δ Female population share	-0.00	0.01	-0.09	0.12
Δ Share of population older than 65	0.03	0.02	-0.12	0.14
Δ Share of population younger than 18	-0.03	0.02	-0.15	0.09
Δ Herfindahl index of age groups	-0.00	0.02	-0.10	0.25
Δ Unemployment rate	-0.01	0.02	-0.10	0.45
Municipality has own district court	0.05	0.21	0.00	1.00
Observations		5544		

Notes: This table shows summary statistics for all variables used in specifications (1.13), (1.14) and (1.15).

municipalities in my sample and equals roughly 0.7% of a standard deviation of that decline. On average, however, the increase in broadband penetration varies by 17 percentage points such that a standard deviation increase causes a 1% of a standard deviation decline of the Herfindahl index. As mentioned before, the first-difference coefficient is likely downward biased, a conjecture confirmed by the negative coefficient on the 2002 Herfindahl index. Municipalities with greater Herfindahl index in 2002, i.e. those with a concentrated vote share distribution were those who experienced significantly greater decreases in concentration over the following seven years. These are the less agglomerated regions in which broadband was rolled out slower. Not surprisingly, the IV coefficient is five times larger than the OLS estimate indicating that a standard deviation increase in broadband penetration actually caused roughly a 5% of a standard deviation increase in the negative difference of the Herfindahl index.

Finally note that, given the high F statistic of 237 in the first stage regression presented in column (4), the instrument is very strong. In the next subsection, I will provide evidence that the exogeneity assumption is also very likely fulfilled which lends credibility to the magnitude and sign of the IV estimate.

Table 1.6: Estimation results, virtual ties

Dependent variable:	Δ Herfindahl index ('09-'02)			Δ Broadband penetration
Specification:	Baseline	IV	Reduced	1st stage
	(1)	(2)	(3)	(4)
Regressors:				
Δ Broadband penetration	-0.004*	-0.020***		
	(0.002)	(0.006)		
Distance to MDF			0.001***	-0.029***
			(0.000)	(0.002)
Herfindahl index 2002	-0.534***	-0.538***	-0.535***	-0.122**
	(0.009)	(0.009)	(0.008)	(0.051)
Δ Population	-0.002**	-0.002**	-0.003**	0.005
	(0.001)	(0.001)	(0.001)	(0.008)
Mean of dependent variable	-0.11	-0.11	-0.11	0.90
Obs.	5544	5544	5544	5544
R^2	0.900		0.900	
Kleibergen-Paap rk Wald F statistic				237
Anderson-Rubin Wald F statistic				12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows the results of specifications (1.13), (1.14) and (1.15) in columns (1), (4) and (2), respectively. Column (3) shows the reduced form regression of the IV specification in column (2). Control variables include differences of the following variables: A municipality's population (as a cubic), the female population share, the share of individuals older than 65, younger than 18, the Herfindahl index of age group shares and the unemployment rate. Additionally included are the Herfindahl index of vote shares in 2002 and a county fixed effect. Standard errors are clustered at the municipality level.

1.7.3. Robustness checks

In this section, I will provide evidence that the exogeneity assumption related to the instrumental variable of the previous regressions does indeed apply. To do so, I estimate the reduced form specification shown in column (3) of Table 1.6 using differences between the 2002 and 1994 Herfindahl index as dependent variable instead:

$$\Delta H_{m,'02-'94} = \beta_0 + \beta_c + \beta_1 D_m + \boldsymbol{\Delta X'_{m,'02-'94}} \boldsymbol{b} + \epsilon_m \tag{1.16}$$

Table 1.F.2 in the appendix shows summary statistics of the variables used in this regression and their respective values in 2002 and 1994. If the distance of a municipality to its MDF is related to the vote share distribution in any other way than through its effect on broadband availability, conditional on controls, one would expect a significant coefficient in such a regression even though broadband technology was barely available

before 2002. Due to the data restrictions already described, I am using 1993 and 2001 control variables and cannot include the unemployment rate in this specification.[35] Additionally, I repeat the estimations of the models shown in columns (1), (2) and (3) of Table 1.6 calculating the Herfindahl index also based on the vote share of the Green party.

Table 1.7 shows the respective results. Models (1) to (3) correspond to the respective models estimated in Table 1.6. The baseline coefficient is increased by 0.001 and therefore loses its significance. The IV and the reduced form estimate of column (3), however, remain constant. Importantly, the distance of a municipality to its MDF has no significant influence on the development of political fragmentation between 1994 to 2002. This holds across both specifications of the Herfindahl index shown in columns (4) and (5), respectively. The findings provide strong evidence for the exclusion restriction of the instrument, i.e. that the distance of a municipality to its MDF, conditional on control variables, only affects political fragmentation through its effects on broadband availability.

[35]Table 1.F.3 in the appendix repeats all analyses presented in Table 1.6 excluding the unemployment rate as control variable. Since point estimates are almost identical, it is unlikely that excluding the unemployment rate from the placebo regressions will affect the results.

Table 1.7: Estimation results, virtual ties, robustness checks

Herfindahl index based on:	CDU, SPD, FDP and Green party				CDU, SPD and FDP
Dependent variable:	Δ Herfindahl index ('09-'02)			Δ Herfindahl index ('02-'94)	
Specification:	Baseline	IV	Reduced	Reduced (placebo)	Reduced (placebo)
	(1)	(2)	(3)	(4)	(5)
Regressors:					
Δ Broadband penetration	-0.003	-0.021***			
	(0.002)	(0.007)			
Distance to MDF			0.001***	0.000	0.000
			(0.000)	(0.000)	(0.000)
Herfindahl index 2002	-0.474***	-0.478***	-0.476***		
	(0.009)	(0.009)	(0.009)		
Herfindahl index 1994				-0.189***	-0.183***
				(0.012)	(0.012)
Δ Population	-0.004***	-0.004***	-0.004***	-0.012***	-0.012***
	(0.001)	(0.001)	(0.001)	(0.008)	(0.001)
Mean of dependent variable	-0.13	-0.13	-0.13	0.01	0.01
Obs.	5544	5544	5544	5638	5638
R^2	0.873		0.873	0.813	0.810

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows the results of specifications (1.13) and (1.15) in columns (1) and (2) respectively. Column (3) shows the reduced form regressions of the IV specification in column (2). Columns (4) and (5) show results of a regression similar to that of column (3) but use differences between the 2002 and 1994 Herindahl index as dependent variable. For columns (1) to (4) the Herfindahl index is calculated based on vote shares of the CDU, SPD, FDP and the Green party. In column (5) it is based only on the first three of the aforementioned parties. Controls include the Herfindahl index of vote shares in 2002, the difference in a municipality's population (as a cubic), female population share, share of individuals younger than 18, individuals older then 65, Herfindahl index of age groups, unemployment rate (not for models of columns (4) and (5)) and a county fixed effect. Standard errors are clustered at the municipality level.

1.8. Conclusion

This paper proposes a model which links political fragmentation in a region to the number of social ties between its inhabitants and to the direction of the respective tie formation process. Individuals are embedded in a social network and take into account past partisan affiliation of their peers when maximizing their own utility with respect to their voting decisions. I show that vote share distributions disperse when the number of social ties in the network declines or when ties are more likely to emerge between voters of the same party.

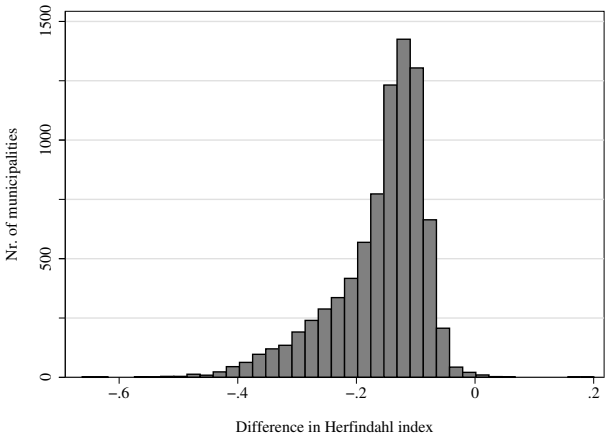
I use this model to relate the decline in face-to-face and the subsequent increase in virtual interactions between German citizens to the dispersion of federal election vote shares within one comprehensive framework. First, I provide evidence for the existence of peer effects in voting decisions by showing that partisan support of individuals who change residential locations is 55% of a standard deviation higher for parties which constitute the majority in the destination region. I also find no evidence that regional sorting could explain this result. Second, using the number of clubs and associations in a region as proxy for face-to-face interactions, my IV estimates suggest that a standard deviation decrease in this variable causes about 16% of a standard deviation decrease in the Herfindahl index. Finally, building on the intuition that virtual ties are much more likely to form between individuals with the same partisan affiliation, I find that a standard deviation increase in the percentage of households connected to the broadband internet caused roughly 5% of a standard deviation increase in the decline of the Herfindahl index between 2002 and 2009.

My results bear implications for different fields. There is evidence that fragmentation of governments has a significant impact on budget size (Ricciuti, 2004; Elgie and McMenamin, 2008; Eslava and Nupia, 2010; Kim and Kim, 2021) and voter turnout (Zagórski, 2021). Furthermore, many studies argue that fragmented and polarized societies might be a danger for democracy or at least detrimental to the democratic discourse (Sunstein, 2001; Montalvo and Reynal-Querol, 2005; Olken, 2009). A better understanding of how peer effects shape political fragmentation can help to mitigate these dangers. My research also sheds light on how politicians might abuse social influence effects to increase vote shares or stabilize a status-quo vote share distribution (Acemoglu, Reed, and Robinson, 2014; Satyanath, Voigtländer, and Voth, 2017). An interesting avenue for further research would be to quantify the contribution of declines in face-to-face and increases in virtual ties to overall fragmentation or to explore in depth the effects of online social networks like Facebook and Twitter on real-life political outcomes.

Appendix

1.A. Appendix to Section 1.1

Figure 1.A.1: Distribution of differences between 2017 and 1980 Herfindahl index



Notes: This histogram plots the empirical distribution of the difference between the 2017 and 1980 Herfindahl index of all 8245 West German municipalities in my sample. 1980 values are subtracted from 2017 values.

Table 1.A.1: Overview of German parties

Abbreviation *Vote share 2021*	Name *in German*	Orientation/Label
Current		
SPD *25,7%*	Social Democratic Party of Germany *Sozialdemokratische Partei Deutschlands*	Social Democratic
CDU *18,9%*	Christian Democratic Union of Germany *Christlich Demokratische Union Deutschlands*	Conservative
CSU *5,5%*	Christian Social Union in Bavaria[1] *Christlich-Soziale Union in Bayern*	Conservative
The Greens *14,8%*	Alliance 90/The Greens *Bündnis 90/Die Grünen*	Green
AfD *10,3%*	Alternative for Germany *Alternative für Deutschland*	Right
FDP *11,5%*	Free Democratic Party *Freie Demokraten*	Liberal
The Left *4,9%*	The Left *Die Linke*	Left
Former		
SED	Socialist Unity Party of Germany *Sozialistische Einheitspartei Deutschlands*	Socialist
PDS	Party of Democratic Socialism *Partei des Demokratischen Sozialismus*	Left/Socialist
WASG	Labour and Social Justice – The Electoral Alternative *Arbeit & soziale Gerechtigkeit – Die Wahlalternative*	Left

Notes: This table provides an overview of relevant political parties in Germany and their current shares of second votes cast in the 2021 federal election.
1: The CDU is eligible for elections in all German federal states except of Bavaria where, instead, the CSU competes for votes. In federal parliament, the parties form the so-called *Union* and are part of one parliamentary group.

1.A.1. German electoral system and postwar federal elections

Electoral system In German federal elections each citizen has two votes. Half of the parliament's seats are allocated to party representatives who won the simple majority (that is highest share) of first votes within their electoral district. The remaining seats are distributed across parties such that the total relative proportions mirror the outcome of second votes cast. If a party wins so many electoral districts by first votes that its share in parliament is higher than that implied by its second vote share, the parliament is enlarged accordingly. Therefore, the second vote is considered the more important one. Note also that due to their first-past-the-post character, it is more likely that first votes are affected by strategic voting. Individuals might not want to "waste" their first vote on their preferred candidate if it is relatively probable that she will not win the

simple majority. Only parties which receive at least 5% of second votes cast move into parliament unless they are able to win at least three electoral districts by first votes. All results in the main text are based on second votes.

Postwar federal elections Since the end of the second world war, German governments have usually been formed by two parties. Either of the large centrist parties, the Christian-Conservative CDU/CSU or the Social Democrats, SPD, formed a coalition with either of the smaller parties, the Liberal party, FDP, or the Green party.[36] However, in 2005, vote shares were so dispersed that neither of these coalitions represented a majority of the votes. This trend continued, interrupted by a coalition between CDU and FDP from 2009 to 2013, with federal elections in 2013, 2017 and 2021. In 2013 and 2017, CDU and SPD had to form a *grand coalition* to ensure a majority in parliament. 2021 marked the first year in which a postwar German government consisted of three parties.

After the German reunification in 1990, two parties emerged whose successor organizations are still relevant for today's political landscape in Germany. First, in 1990, the PDS, successor of the SED which had ruled the German Democratic Republic. Second, in 1993, the Green party (Alliance 90/The Greens) which was founded by a merger of the West German Green party and the East German civil rights movement Alliance 90.

In 2007, after previous collaboration already for federal elections in 2005, the left wing parties PDS and WASG officially joined forces to become the Left party, *Die Linke*. Importantly, since the WASG had not been eligible for federal elections, the number of parties in parliament was not affected by this merger. However, together as *Die Linke*, the parties were able to significantly increase their vote share and thereby disperse the political spectrum.

In 2013, right wing party AfD was eligible for federal elections for the first time and managed to enter parliament in 2017.

[36] An exception being the period from 1966 to 1969 in which the two centrist parties formed a so-called *grand coalition*.

Table 1.A.2: Dispersing opinions in Germany

Topic/Question	Development of share of answer options selected and Herfindahl index			
European integration should proceed[1]	**1982**	**1992**	**2002**	**2011**
Faster	0.50	0.13	0.25	0.14
Slower	0.06	0.38	0.22	0.43
As is	0.28	0.36	0.41	0.30
Herfindahl index	**0.36**	**0.31**	**0.29**	**0.31**
Technological progress will make life[2]	**1966**	**1981**	**2011**	
Easier	0.50	0.32	0.40	
More difficult	0.29	0.44	0.35	
The same	0.11	0.16	0.17	
Herfindahl index	**0.36**	**0.33**	**0.32**	
Conflict between Islam and Western culture[3]	**2006**	**2010**	**2012**	
Will occur	0.55	0.48	0.44	
Exists already	0.22	0.21	0.25	
Will not occur	0.14	0.21	0.21	
Herfindahl index	**0.38**	**0.33**	**0.31**	
Task sharing in family should be as follows[4]	**2007**	**2014**	**2019**	
Men works full time, woman part time and mostly takes care of household and children	0.43	0.38	0.36	
Both work part time and take care of household and children	0.19	0.28	0.22	
Men works full time, woman only takes care of household and children	0.20	0.17	0.18	
Both work full time and take care of household and children	0.15	0.10	0.16	
Woman works full time, man part time and mostly takes care of household and children	0.01	0.01	0.02	
Woman works full time, man only takes care of household and children	0.03	0.02	0.01	
Herfindahl index	**0.28**	**0.27**	**0.24**	
Political system[5]	**1988**	**1994**		
Is working well and needs no change	0.18	0.07		
Generally works well but needs some changes	0.66	0.57		
Does not work well and needs many changes	0.15	0.32		
Does not work at all and needs to be changed completely	0.01	0.04		
Herfindahl index	**0.49**	**0.44**		

Notes: This table shows survey results for representative (sub)samples of the German population. Answer option "undecided" is excluded for each question.
1: Source: Institut für Demoskopie (IfD) Allensbach, "Das gemeinsame Interesse an Europa ist in Gefahr". Published in Frankfurter Allgemeine Zeitung (FAZ), Nr. 21, 26.01.2011, p. 5. West German population older than 16. **2:** Source: IfD Allensbach, "Kein Fortschrittspessimismus". Published in FAZ, Nr. 115, 18.05.2011, p.5. West German population older than 16. **3:** Source: IfD Allensbach, "Die Furcht vor dem Morgenland im Abendland". Published in FAZ, Nr. 242, 21.11.2012, p.10. German population older than 16. **4:** Source: IfD Allensbach, "Elternschaft heute". 20.01.2020. German parents of underage children. **5:** Source: "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (Allbus)". West German population.

Table 1.A.3: Representative data on development of social interactions in West Germany

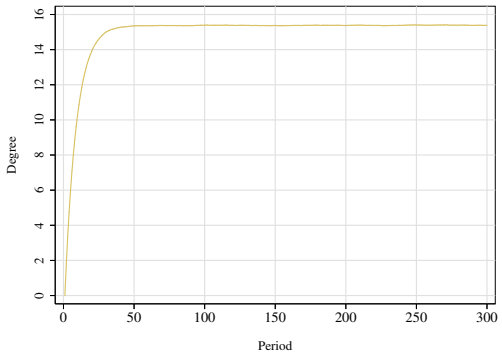
Variable	Scale $[min - max]$ $(Unit)$	Early value	Late value
Face-to-face interaction/Physical social ties			
Member of at least one club/association[12]	Binary	**1980**	**1996**
German population		0.48	0.46
Individuals younger than median age		0.53	0.49
older than median age		0.44	0.43
Average daily time spent on: social contacts[34]	Cardinal *(minutes)*	**2001**	**2012**
German population		40	36
Individuals aged 10 - 17		34	37
18 - 29		52	40
30 - 44		42	33
45 - 64		38	34
older than 65		36	38
Virtual interaction/Virtual social ties			
Free time: use internet or specific online services[2]	Ordinal $[1 = daily - 5 = never]$	**1998**	**2004**
German population		4.64	3.51
Individuals younger than median age		4.41	2.89
older than median age		4.85	4.19
Average daily time spent on: communicating via Computer and Smartphone[4]	Cardinal *(minutes)*	**2001**	**2012**
German population:		1	5
Individuals aged 10 - 17		1	7
18 - 29		3	8
30 - 44		2	5
45 - 64		1	5
older than 65		1	4

Notes: This table shows the development of some indicators of social contacts for representative samples of West German individuals based on two different surveys. **1:** Religious, singing, sports, hobby, homeland, welfare, youth/student, civil initiative and other. **2:** Source: "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (Allbus)". West German population. **3:** Face-to-face contact or phone calls. No internet. **4:** Source: "Wie die Zeit vergeht - Analysen zur Zeitverwendung in Deutschland". Federal Statistical Office of Germany. West German population.

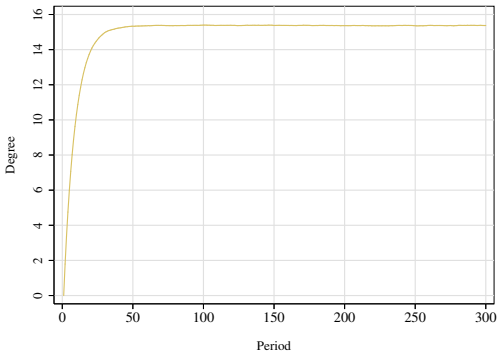
1.B. Appendix to Section 1.3

Figure 1.B.1: Trajectories of degrees for simulations depicted in Figure 1.2

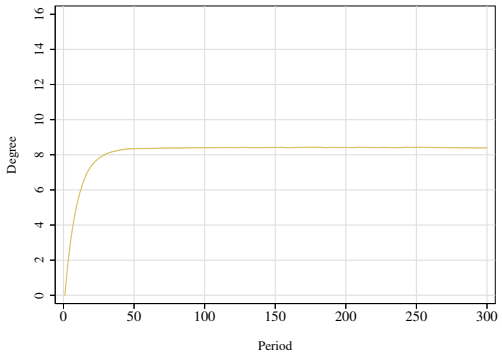
(a) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 1.5$



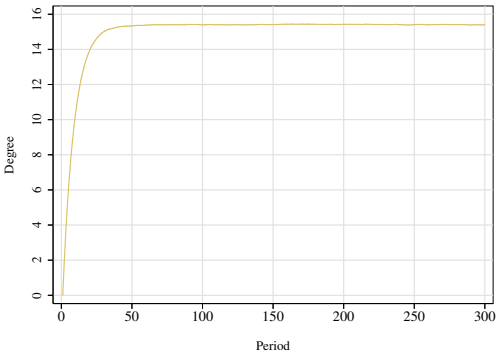
(b) ***Parameters:***
$c_f = 0.01$, $c_B = 0.05$, $R = 1.5$,
$\delta = 0.5$



(c) ***Parameters:***
$\delta = 0.2$, $c_B = 0.05$, $R = 1.5$,
$c_f = 0.005$



(d) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$,
$R = 2$



Notes: This figure shows the average trajectories of degrees across 50 simulation loops of the model presented in Section 1.3. Note that the plots show averages across simulations of $d^* = \frac{D^*}{N}$ which is itself an average across individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$ and preference parameters s_{ik} follow standard normal distributions which are independent across individuals and $c_B = 0.05$. In each panel, one of the parameters δ, c_F and R is varied as compared to the baseline specification in panel (a).

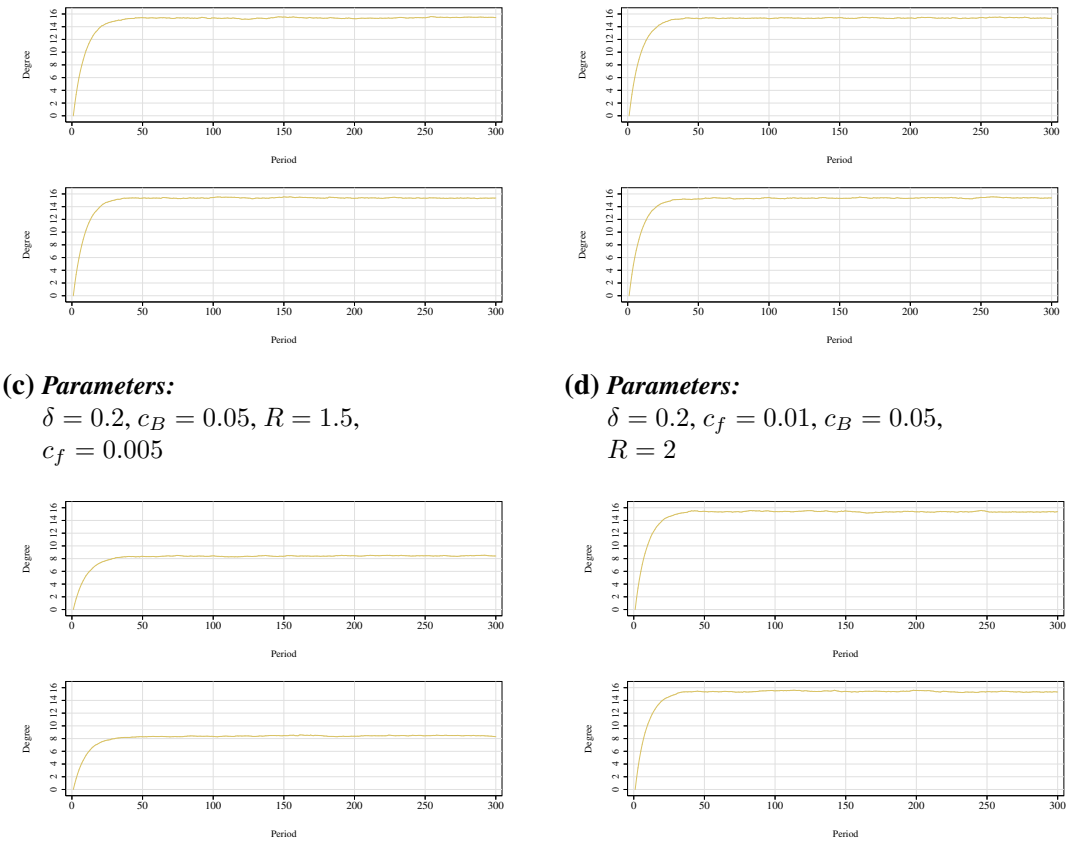
Figure 1.B.2: Exemplary trajectories for simulations depicted in Figure 1.B.1

(a) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 1.5$

(b) ***Parameters:***
$c_f = 0.01$, $c_B = 0.05$, $R = 1.5$, $\delta = 0.5$

(c) ***Parameters:***
$\delta = 0.2$, $c_B = 0.05$, $R = 1.5$, $c_f = 0.005$

(d) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 2$



Notes: This figure shows exemplary trajectories of degrees based on one simulation loop of the model presented in Section 1.3. Note that the plots show exemplary trajectories of $d^* = \frac{D^*}{N}$ which is the average across individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$ and preference parameters s_{ik} follow standard normal distributions which are independent across individuals and $c_B = 0.05$. In each panel, one of the parameters δ, c_F and R is varied as compared to the baseline specification in panel (a). Per panel, exemplary trajectories of two distinct simulations are shown.

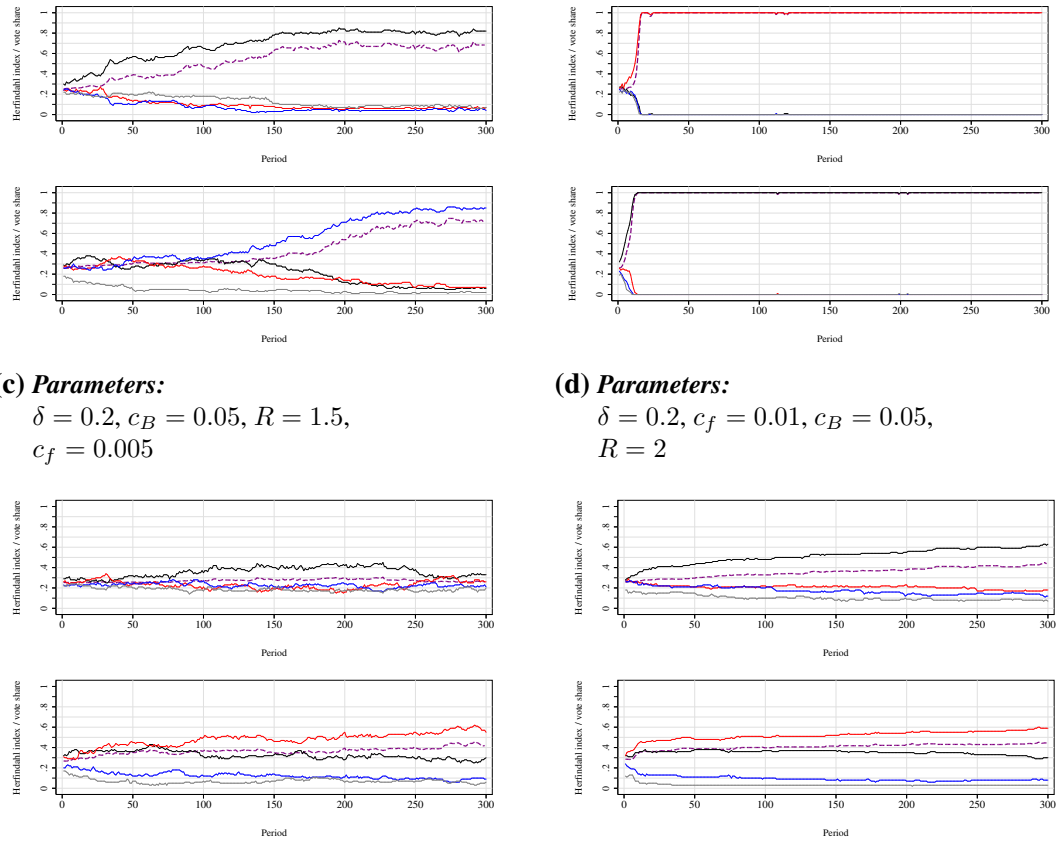
Figure 1.B.3: Exemplary trajectories for simulations depicted in Figure 1.2

(a) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$, $R = 1.5$

(b) ***Parameters:***
$c_f = 0.01$, $c_B = 0.05$, $R = 1.5$,
$\delta = 0.5$

(c) ***Parameters:***
$\delta = 0.2$, $c_B = 0.05$, $R = 1.5$,
$c_f = 0.005$

(d) ***Parameters:***
$\delta = 0.2$, $c_f = 0.01$, $c_B = 0.05$,
$R = 2$

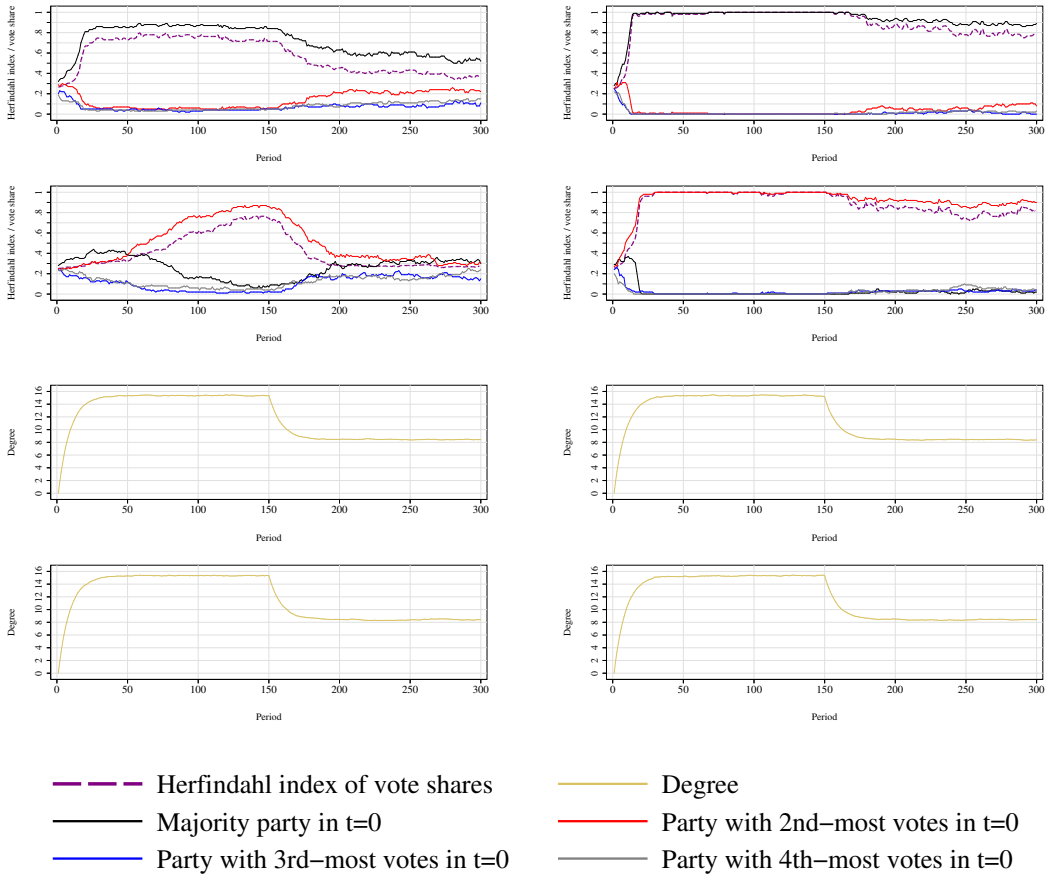


Notes: This figure shows exemplary trajectories of vote shares and the Herfindahl index based on one simulation loop of the model presented in Section 1.3. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals and $c_B = 0.05$. In each panel, one of the parameters δ, c_F and R is varied as compared to the baseline specification in panel (a). Per panel, exemplary trajectories of two distinct simulations are shown.

Figure 1.B.4: Negative shock to c_F: Exemplary trajectories for simulations depicted in Figure 1.3

(a) ***Parameters:***
$c_f = 0.01$ for $t \leq 150$ and
$c_f = 0.005$ for $t > 150$,
$c_B = 0.05$, $R = 1.5$,
$\delta = 0.2$

(b) ***Parameters:***
$c_f = 0.01$ for $t \leq 150$ and
$c_f = 0.005$ for $t > 150$,
$c_B = 0.05$, $R = 1.5$,
$\delta = 0.5$

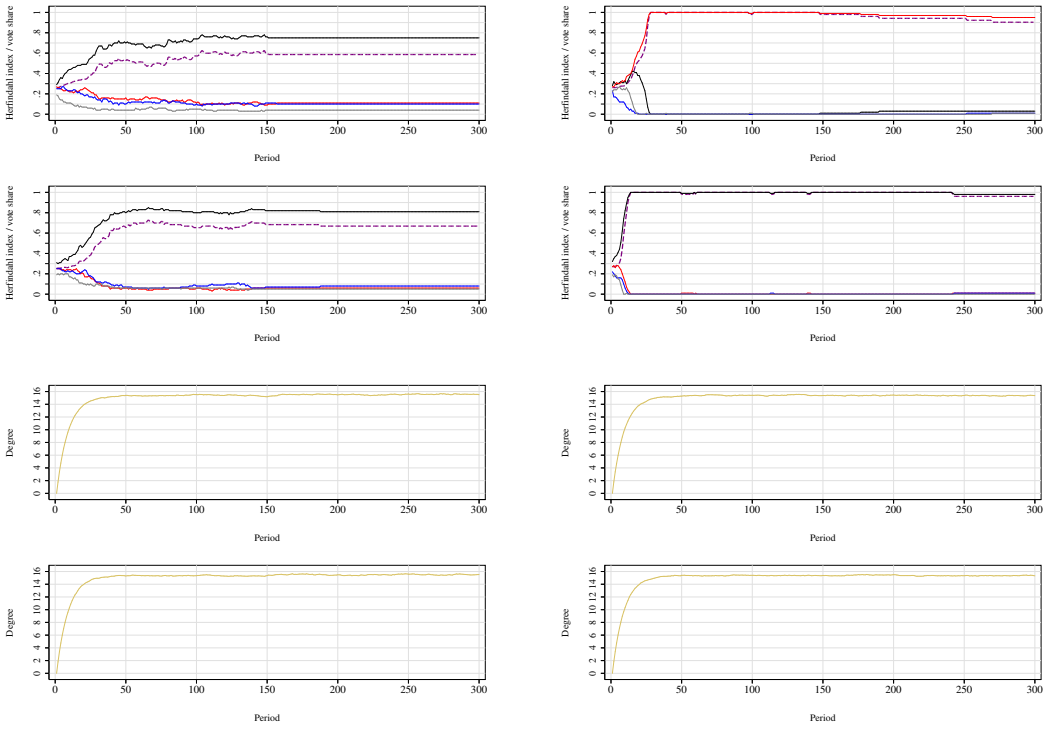


Notes: This figure shows exemplary trajectories of vote shares, the Herfindahl index and degrees based on one simulation loop of the model presented in Section 1.3. Note that the lower panels plot exemplary trajectories of $d^* = \frac{D^*}{N}$ which is the average actoss individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals, $c_B = 0.05$, $R = 1.5$ and $c_F = 0.01$ for $t \leq 150$ and $c_F = 0.005$ for $t > 150$. $\delta = 0.2$ in panel (a) and $\delta = 0.5$ in panel (b). Per panel, exemplary trajectories of two distinct simulations are shown.

Figure 1.B.5: Positive shock to R: Exemplary trajectories for simulations depicted in Figure 1.4

(a) ***Parameters:***
$c_F = 0.01$, $c_B = 0.05$,
$R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$,
$\delta = 0.2$

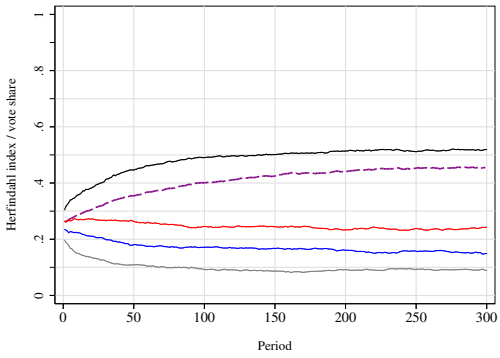
(b) ***Parameters:***
$c_F = 0.01$, $c_B = 0.05$,
$R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$,
$\delta = 0.5$



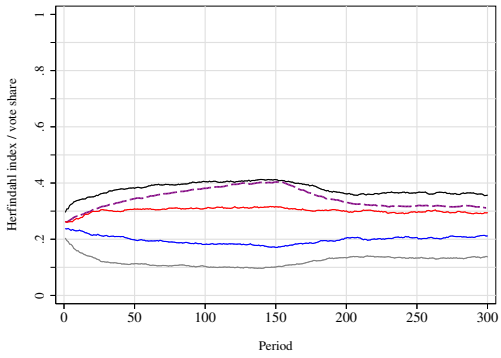
Notes: This figure shows exemplary trajectories of vote shares, the Herfindahl index and degrees based on one simulation loop of the model presented in Section 1.3. Note that the lower panels plot exemplary trajectories of $d^* = \frac{D^*}{N}$ which is the average actoss individual degrees in the network. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals, $c_F = 0.01$, $c_B = 0.05$ and $R = 1.5$ for $t \leq 150$ and $R = 5$ for $t > 150$. $\delta = 0.2$ in panel (a) and $\delta = 0.5$ in panel (b). Per panel, exemplary trajectories of two distinct simulations are shown.

Figure 1.B.6: Decreasing marginal utility: Simulation results for Herfindahl index and vote shares

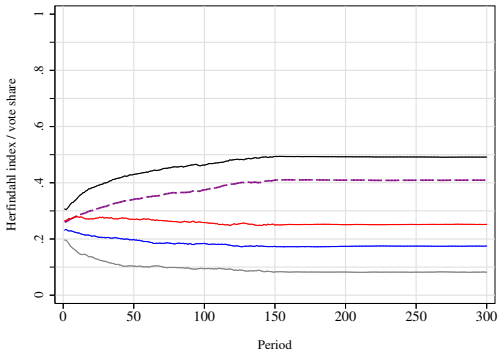
(a) ***Parameters:***
$\delta = 0.65$, $c_f = 0.01$, $c_B = 0.05$, $R = 1.5$

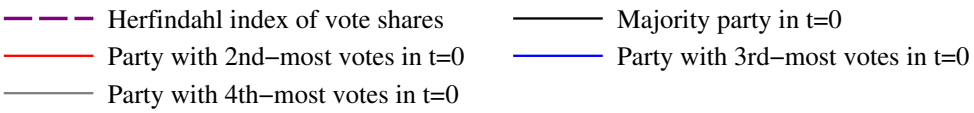


(b) ***Parameters:***
$\delta = 0.65$
$c_f = 0.01$ for $t \leq 150$ and $c_f = 0.005$ for $t > 150$,
$c_B = 0.05$, $R = 1.5$



(c) ***Parameters:***
$\delta = 0.65$, $c_f = 0.005$, $c_B = 0.05$,
$R = 1.5$ for $t \leq 150$ and $R = 2$ for $t > 150$

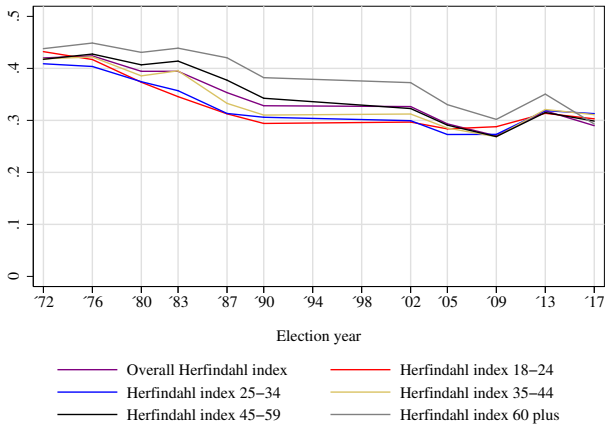




Notes: This figure shows the average trajectories of vote shares and the Herfindahl index based on 50 simulation loops of the model presented in Section 3.1. Across all panels, the number of periods is set to $T = 300$, the number of individuals to $N = 100$, utility is specified as $U_{it}(o_{it}) = \sum_{o_k \in O} \mathbb{1}\{o_{it} = o_k\} s_{ik} + \delta \sqrt{\sum_{j \neq i}^{N} g_{ijt} \mathbb{1}\{o_{it} = o_{jt}\}}$, preference parameters s_{ik} follow standard normal distributions which are independent across individuals, $\delta = 0.65$ and $c_B = 0.05$. In panel (a) $c_f = 0.01$ and $R = 1.5$. In panel (b) $c_f = 0.01$ for $t \leq 150$ and $c_f = 0.005$ for $t > 150$ and $R = 1.5$. In panel (c) $c_f = 0.01$ and $R = 1.5$ for $t \leq 150$ and $R = 2$ for $t > 150$.

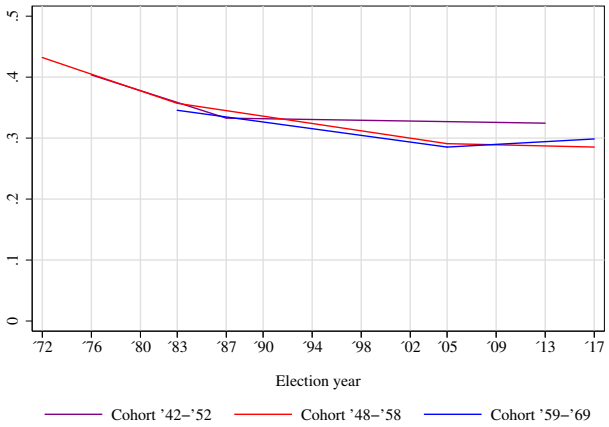
Figure 1.B.7: Herfindahl index by age groups and birth cohorts

Herfindahl index by age groups



Notes: This figure plots the federal level Herfindahl index of vote shares in federal elections since 1972 split by age groups. Election years are not evenly spaced because legislative periods can end before the regular election date. The data was collected in representative surveys conducted on election days and are taken from the website of the Bundeswahlleiter (compare Section 1.4). Data for 1994 and 1998 are missing because surveys were not conducted in these years. Data for 2021 are not available yet.

Herfindahl index by birth cohorts



Notes: This figure plots the federal level Herfindahl index of vote shares in federal elections since 1972 split by birth cohorts. Election years are not evenly spaced because legislative periods can end before the regular election date. This plot is based on data which was collected in representative surveys conducted on election days and is taken from the website of the Bundeswahlleiter (compare Section 1.4). The original data is split by age groups, birth cohorts are recovered from that data. Original data for 1994 and 1998 are missing. Data for 2021 are not available yet.

1.C. Appendix to Section 1.4

Table 1.C.1: Data sources

Variable	Available Years	Source	Scale $[min - max]$ $(Unit)$
Individual level			
Municipality of residence	1984-2014	SOEP	Categorical
Supported party	1984-2014	SOEP	Categorical
Degree of party support	1984-2014	SOEP	Ordinal $[1 - 5]$
Gender	1984-2014	SOEP	Categorical
Migration background	1984-2014	SOEP	Categorical
Regional level			
Municipality			
Federal election vote shares	1980-2017	Bundeswahlleiter	Cardinal
Number of clubs and associations	2019	Vereinsregister	Cardinal
Municipality has own district court	2021	Anwaltsverzeichnis für Deutschland	Categorical
Share of households connected to broadband network (Broadband penetration rate)	1980-2001, 2005-2008	Falck, Gold, and Heblich (2014)	Cardinal
Distance to MDF	1980-2017	Falck, Gold, and Heblich (2014)	Cardinal (km)
Area	2008-2017	German Office for Statistics	Cardinal (km^2)
Population by age and gender	1993, 2001, 2008-2017	German Office for Statistics	Cardinal
Number of unemployed	2008-2017	German Office for Statistics	Cardinal
Business tax multiplier	2008-2017	German Office for Statistics	Cardinal
Business tax revenue	2008-2017	German Office for Statistics	Cardinal (€)
County			
Population by education and nationality	2008-2017	German Office for Statistics	Cardinal

Notes: This table gives an overview of all data used throughout this paper. The Herfindahl index is calculated from federal election vote shares. The female population share, the share of the population older than 65, the share of the population younger than 18 and the Herfindahl index of age group shares are calculated from municipality level population data which is split by gender and age groups. The foreign population share, the Herfindahl index of foreign nationalities and the population share of non-educated are calculated from county level population data which is split by nationalities and education levels.

1.D. Appendix to Section 1.5

The following tables show results of the same regressions as shown in Table 1.2 (Equations (1.8) and (1.9)) but for different subsamples. Tables 1.D.1 to 1.D.3 repeat the analyses only for those individuals who moved after years 2000, 2005 and 2010 respectively.

Table 1.D.1: Estimation results, testing for peer effects, 2000-2014

Dependent variable:	Party support		Elected same party
Pre- / Post-move:	Post	Pre	Post
	(1)	(2)	(3)
Regressors:			
Majority=1	0.634***	-0.009	0.227***
	(0.090)	(0.015)	(0.026)
Majority=1×Clubs&Assoc.	0.059***	-0.005	0.017**
	(0.020)	(0.005)	(0.007)
Mean of dependent variable	2.99	3.20	0.89
Obs.	1967	3420	1967
R^2	0.444	0.885	0.370
# of groups	382	483	382

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression Equations (1.8) and (1.9) in columns (1) and (2), respectively. Column (3) shows results of a regression similar to that of column (1) but uses a dummy indicating whether an individual voted for the same party before and after moving as dependent variable. Control variables include the destination municipality's size, its population (as a cubic), its population shares of individuals older than 65, younger than 18, of foreigners, of non-educated, its Herfindahl index of foreigner shares, its business tax multiplier and revenue, its unemployment rate, individual gender and migration background. For the post-move regressions, controls additionally include individual party support in the origin region. Each regression also includes the following, interacted fixed effects: An indicator for the supported party before moving, an indicator for different levels of party support and an indicator for different levels of the vote share of the supported party in the origin region. Number of groups indicates into how many groups the sample is partitioned by the interacted fixed effects. Standard errors are clustered at the municipality level. Only individuals who moved between 2000 and 2014 are included in the sample.

Table 1.D.2: Estimation results, testing for peer effects, 2005-2014

Dependent variable:	Party support		Elected same party
Pre- / Post-move:	Post	Pre	Post
	(1)	(2)	(3)
Regressors:			
Majority=1	0.726***	-0.006	0.243***
	(0.122)	(0.018)	(0.037)
Majority=1×Clubs&Assoc.	0.058***	-0.007	0.015**
	(0.020)	(0.005)	(0.007)
Mean of dependent variable	3.01	3.18	0.89
Obs.	1414	2768	1414
R^2	0.512	0.886	0.414
# of groups	316	436	316

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression Equations (1.8) and (1.9) in columns (1) and (2), respectively. Column (3) shows results of a regression similar to that of column (1) but uses a dummy indicating whether an individual voted for the same party before and after moving as dependent variable. Control variables include the destination municipality's size, its population (as a cubic), its population shares of individuals older than 65, younger than 18, of foreigners, of non-educated, its Herfindahl index of foreigner shares, its business tax multiplier and revenue, its unemployment rate, individual gender and migration background. For the post-move regressions, controls additionally include individual party support in the origin region. Each regression also includes the following, interacted fixed effects: An indicator for the supported party before moving, an indicator for different levels of party support and an indicator for different levels of the vote share of the supported party in the origin region. Number of groups indicates into how many groups the sample is partitioned by the interacted fixed effects. Standard errors are clustered at the municipality level. Only individuals who moved between 2005 and 2014 are included in the sample.

Table 1.D.3: Estimation results, testing for peer effects, 2010-2014

Dependent variable:	Party support		Elected same party
Pre- / Post-move:	Post	Pre	Post
	(1)	(2)	(3)
Regressors:			
Majority=1	1.047***	0.032	0.339***
	(0.373)	(0.035)	(0.110)
Majority=1×Clubs&Assoc.	0.086*	-0.008	0.024**
	(0.046)	(0.011)	(0.011)
Mean of dependent variable	3.21	3.22	0.94
Obs.	316	1140	316
R^2	0.631	0.909	0.578
# of groups	108	266	108

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression Equations (1.8) and (1.9) in columns (1) and (2), respectively. Column (3) shows results of a regression similar to that of column (1) but uses a dummy indicating whether an individual voted for the same party before and after moving as dependent variable. Control variables include the destination municipality's size, its population (as a cubic), its population shares of individuals older than 65, younger than 18, of foreigners, of non-educated, its Herfindahl index of foreigner shares, its business tax multiplier and revenue, its unemployment rate, individual gender and migration background. For the post-move regressions, controls additionally include individual party support in the origin region. Each regression also includes the following, interacted fixed effects: An indicator for the supported party before moving, an indicator for different levels of party support and an indicator for different levels of the vote share of the supported party in the origin region. Number of groups indicates into how many groups the sample is partitioned by the interacted fixed effects. Standard errors are clustered at the municipality level. Only individuals who moved between 2010 and 2014 are included in the sample.

1.E. Appendix to Section 1.6

Table 1.E.1: Estimation results, physical ties, alternative specification of Herfindahl index

Dependent variable:	Herfindahl index 2017			Clubs&Assoc.
Specification:	Baseline	IV	Reduced	1st stage
	(1)	(2)	(3)	(4)
Regressors:				
Clubs&Assoc.	0.005	0.091***		
	(0.010)	(0.035)		
Municipality has district court			0.003***	0.035***
			(0.001)	(0.003)
Population	-0.003***	-0.003***	-0.003***	0.005***
	(0.000)	(0.000)	(0.000)	(0.001)
Unemployment rate	-0.241***	-0.255***	-0.246***	0.101***
	(0.039)	(0.041)	(0.040)	(0.037)
Mean of dependent variable	0.28	0.28	0.28	0.03
Obs.	7791	7791	7791	7791
R^2	0.552		0.552	
Kleibergen-Paap rk Wald F statistic				128
Anderson-Rubin Wald F statistic				7

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regressions Equations (1.10), (1.11) and (1.12) in columns (1), (4) and (2), respectively. Column (3) shows the reduced form regression of the IV specification in column (2). Control variables include a municipality's size, its population (as a cubic), the female population share, the share of individuals older than 65, younger than 18, its Herfindahl index of age group shares, its business tax multiplier and revenue, its unemployment rate, the distance to its MDF and a county fixed effect. Standard errors are clustered at the municipality level. The dependent variable, the Herfindahl index of vote shares in 2017 is additionally based on the vote share of the Green party.

1.F. Appendix to Section 1.7

Table 1.F.1: Summary statistics, virtual ties

Variable	Mean	Std.Dev.	Min.	Max.
2009				
Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.30	0.04	0.25	0.62
Based on CDU, SPD, FDP and Green party	0.28	0.05	0.21	0.62
Share of households connected to broadband network	0.90	0.17	0.00	1.00
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Population (*in 1000*)	5.41	10.52	0.10	151.28
Female population share	0.50	0.02	0.12	0.60
Share of population older than 65	0.20	0.04	0.08	0.42
Share of population younger than 18	0.18	0.03	0.04	0.35
Herfindahl index of age-groups	0.46	0.02	0.36	0.68
Unemployment rate	0.04	0.02	0.00	0.49
Municipality has own district court	0.05	0.21	0.00	1.00
2002				
Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.42	0.09	0.26	0.80
Based on CDU, SPD, FDP and Green party	0.41	0.09	0.23	0.80
Share of households connected to broadband network	0.00	0.00	0.00	0.00
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Population (*in 1000*)	5.43	10.61	0.10	150.96
Female population share	0.50	0.02	0.14	0.61
Share of population older than 65	0.17	0.03	0.06	0.40
Share of population younger than 18	0.21	0.03	0.07	0.39
Herfindahl index of age-groups	0.46	0.02	0.36	0.65
Unemployment rate	0.04	0.02	0.00	0.16
Municipality has own district court	0.05	0.21	0.00	1.00
Observations		5544		

Notes: This table shows summary statistics for the 2009 (2008) and 2002 (2001) values of the variables used in specifications (1.13), (1.14) and (1.15). The Herfindahl indices are based on 2002 and 2009 values, all other variables are based on 2008 and 2001 values, respectively except of the distance of a municipality to its MDF and whether a municipality has a district court, which are time-invariant.

Table 1.F.2: Summary statistics, virtual ties

Variable	Mean	Std.Dev.	Min.	Max.
Δ2002-1994				
Δ Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.01	0.06	-0.23	0.24
Based on CDU, SPD, FDP and Green party	0.01	0.07	-0.24	0.26
Share of municipalities with decreasing Herfindahl index	0.51	0.50	0.00	1.00
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Δ Population (*in 1000*)	0.26	0.60	-4.60	10.74
Δ Female population share	-0.00	0.01	-0.09	0.11
Δ Share of population older than 65	0.02	0.02	-0.17	0.18
Δ Share of population younger than 18	-0.00	0.02	-0.21	0.17
Δ Herfindahl index of age-groups	-0.02	0.02	-0.15	0.15
Municipality has own district court	0.05	0.21	0.00	1.00
2002				
Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.42	0.09	0.26	0.80
Based on CDU, SPD, FDP and Green party	0.41	0.09	0.23	0.80
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Population (*in 1000*)	5.40	10.62	0.10	150.96
Female population share	0.50	0.02	0.14	0.63
Share of population older than 65	0.17	0.03	0.06	0.40
Share of population younger than 18	0.21	0.03	0.07	0.39
Herfindahl index of age-groups	0.46	0.02	0.35	0.65
Municipality has own district court	0.05	0.21	0.00	1.00
1994				
Herfindahl index of vote shares				
Based on CDU, SPD and FDP	0.41	0.06	0.27	0.74
Based on CDU, SPD, FDP and Green party	0.40	0.06	0.25	0.74
Distance to MDF (*in km*)	2.92	1.77	0.04	14.83
Population (*in 1000*)	5.15	10.35	0.10	148.56
Female population share	0.50	0.02	0.17	0.68
Share of population older than 65	0.15	0.03	0.05	0.39
Share of population younger than 18	0.21	0.03	0.08	0.37
Herfindahl index of age-groups	0.48	0.03	0.37	0.62
Municipality has own district court	0.05	0.21	0.00	1.00
Observations		5563		

Notes: This table shows summary statistics for all variables used in specification (1.16) and for their respective values in 2002 and 1994. The Herfindahl indices are based on (differences between) 1994 and 2002 values, all other variables are based on (differences between) 2001 and 1993 values, respectively except of the distance of a municipality to its MDF and whether a municipality has a district court, which are time-invariant.

Table 1.F.3: Estimation results, virtual ties, excluding unemployment rate as control variable

Dependent variable:	Δ Herfindahl index ('09-'02)			Δ Broadband penetration
Specification:	Baseline	IV	Reduced	1st stage
	(1)	(2)	(3)	(4)
Regressors:				
Δ Broadband penetration	-0.004*	-0.020***		
	(0.002)	(0.006)		
Distance to MDF			0.001***	-0.029***
			(0.000)	(0.002)
Herfindahl index 2002	-0.534***	-0.538***	-0.535***	-0.122**
	(0.009)	(0.009)	(0.008)	(0.051)
Δ Population	-0.002**	-0.002**	-0.003**	0.005
	(0.001)	(0.001)	(0.001)	(0.008)
Mean of dependent variable	-0.11	-0.11	-0.11	0.90
Obs.	5544	5544	5544	5544
R^2	0.899		0.899	
Kleibergen-Paap rk Wald F statistic				237
Anderson-Rubin Wald F statistic				12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows the results of specifications (1.13), (1.14) and (1.15) in columns (1), (4) and (2), respectively. Column (3) shows the reduced form regression of the IV specification in column (2). Control variables include differences of the following variables: A municipality's population (as a cubic), the female population share, the share of individuals older than 65, younger than 18 and the Herfindahl index of age group shares. Additionally included are the Herfindahl index of vote shares in 2002 and a county fixed effect. Standard errors are clustered at the municipality level.

Chapter 2

THE EFFECT OF BROADBAND INTERNET ON IMMIGRANT INTEGRATION IN GERMANY

2.1. Introduction

This paper studies how using the broadband internet impacts the integration of immigrants into the German society. Integration, according to the German Residence Act (*Aufenthaltsgesetz*) is a key requirement for enabling immigrants to independently partake in economic, social and cultural everyday life. This goal is to be achieved by fostering labor market participation and familiarizing foreign-born individuals with the German culture, history and language.[1] In 2015, the so-called refugee crisis fueled an increasingly critical debate about international migration in Germany. For example, the right-wing party AfD was able to more than double its vote share between federal elections in 2013 and 2017. By many, integration is viewed as an effective tool to alleviate the problems allegedly caused by international migration. This view found its way into legislation when the German government decided on a reform of asylum laws, the so-called *Asylpakete* in September 2015. Among new rules for the allocation of refugees into German municipalities, the distribution of dedicated funds between federal states,

[1] Another definition of the term integration is given, for example, in Berry (2004) according to which it is one of four strategies individuals apply when undergoing the "process of cultural and psychological change resulting from contact between cultural groups and their individual members." Integration describes the strategy where immigrants seek to participate in the host country society while also maintaining their heritage culture. It poses a compromise between the strategy of *Assimilation*, participation in the host country society without maintaining one's own cultural heritage, and *Separation*, the wish to maintain only one's own cultural heritage. The forth strategy is named *Marginalization* which describes the process where immigrants lose touch with their own cultural heritage and neither have the wish to participate in the host country's society. Here I will use the term integration as implied by the German Residence Act.

and re-classifications regarding safe countries of origin, a key element was an increased budget for integration courses.

The scientific community has long recognized the importance of the internet for immigrants and its influence on their integration success. Most of the relevant studies emphasize how migrants use the internet to consume mass media or communicate with family and friends from their respective home countries (compare, amongst others, Melkote and Liu, 2000 or Yoon, 2017). Using the internet in this way is found to impact ethnic identities, that is, for example, the adoption of new values and behaviors or the extent to which individuals identify as citizens of the host country. Another common result is that the possibility to stay in contact with their origin countries increases migrants' life satisfaction and their willingness to remain in the host country permanently. However, different strands of literature, especially on peer and mass media effects, imply that better communication possibilities and mass media exposure, facilitated by the internet, are likely to affect a much wider spectrum of characteristics important for integration (e.g. Chong and Ferrara, 2009 or Brekke and Brochmann, 2015). Furthermore, many studies investigating the link between internet exposure and immigrant integration are based on small samples (often less than 100 individuals) and suffer from identification issues such that the found relationships are only of correlational nature (e.g. Melkote and Liu, 2000 or Yoon, 2017). This paper is intended to fill this gap and provide a comprehensive and statistically sound analysis of the effects of the internet on the various dimensions of integration.

I observe a panel of foreign-born individuals in Germany covering periods well before and after the introduction of broadband technology.[2] The data cover a wide range of relevant integration dimensions. These include measures of ethnic identity, life satisfaction, and willingness to remain in Germany permanently but also of labor market participation, language skills and social contacts with Germans. Importantly, in 2008, all individuals in the sample were asked whether their households were equipped with a broadband access and about the frequency with which they use the internet. These information allow me to construct two indicators of broadband exposure and study their relationship with integration outcomes. Intuitively, the frequency with which an immigrant uses the broadband internet corresponds to the intensive margin of broadband exposure whereas the existence of a broadband connection is a measure for the extensive margin.

The results indicate that individuals with a broadband internet access are 5 percentage points more likely to be employed, 8 percentage points more likely to speak German

[2] In Germany, effective and especially fast use of internet services was only possible with a broadband connection. Digital Subscriber Line (DSL) was the main technology to access the broadband network in Germany for the period I am investigating.

at home and perform voluntary activities in their local communities roughly 16% of a standard deviation more frequently than migrants without a broadband access. Those who use the broadband internet relatively more frequently perform voluntary activities 20% of a standard deviation more frequently compared to infrequent broadband users. Unlike suggested in the literature, I cannot find any causal effect on ethnic identities, willingness to remain in Germany permanently or life satisfaction.

To identify these effects of broadband exposure on integration, I first exploit the panel structure of my data and regress individual changes in integration outcomes of foreign-born individuals between the pre and post-broadband period on the respective changes of the broadband exposure variables and a set of adequate control variables and fixed-effects. These first-difference OLS regressions constitute the baseline specifications. Since some sources of endogeneity such as reverse causality might not be sufficiently controlled for by the baseline specifications, I use a dataset constructed by Falck, Gold, and Heblich (2014) which contains an instrumental variable predicting the so-called broadband penetration rate for municipalities in Germany. The broadband penetration rate measures the share of households connected to the broadband network within a municipality. In their paper, the authors implement a two-stage least squares estimation to investigate the effect of regional broadband availability on voter turnout. The instrument is based on the distance of a municipality to dedicated infrastructure crucial for the provision of broadband services at the early stage of this technology in Germany. Inhabitants of municipalities located further away from the necessary infrastructure were not able to access the broadband internet due to too low bandwidth availability. The instrument is likely exogeneous to the integration success of immigrants because when the necessary infrastructure was built, well before the invention of the internet, it was not known that its location or even its existence would become essential for broadband provision. The most important determinant for location choices was the availability of free lots which, conditional on control variables, were not placed in regions exhibiting characteristics detrimental or beneficial for the integration of the respective immigrant population. Consequently, the instrument should allow me to provide causal evidence for the effect of broadband availability on migrants' integration outcomes. However, I cannot implement a two-stage least squares regression because the instrument, while sufficiently strong for predicting local broadband penetration rates, is neither significantly correlated with household level broadband access nor the frequency of broadband use. Most likely, individuals, especially foreign-born, did not know about the relevance of their residential location for internet bandwidth and thus did not consider it when deciding to purchase a broadband access for their households. Therefore, I have to resort to two different approaches: first, I estimate reduced

form models of the baseline specifications in which I substitute the broadband exposure indicators with the instrumental variable. However, given the weak predictive power of the instrument, I opt for a second, placebo-type approach which constitutes my preferred specification: for each integration indicator, I repeat the baseline specifications on two different samples. One consists of migrants living in areas with high broadband availability, the other is comprised of those individuals located so far away from the necessary infrastructure that they are unable to access the broadband internet, irrespective of the values of the broadband exposure measures. If the baseline results are not driven by biases and the exogeneity assumption of the instrument holds, coefficients should lose significance for the sample of migrants living in municipalities with low broadband availability and remain or become even more significant in the other sample.

A major concern regarding the validity of the exogeneity assumption is that some types of endogeneity might only distort estimates in the sample of individuals living in areas with good broadband connectivity. To rule out this possibility, I perform another type of placebo analysis using time variation. I repeat the baseline regressions on the same sample of individuals, however using changes in the relevant variables well before the pre-broadband period. The coefficients of this robustness check indicate that results of the preferred specifications are unbiased.

According to the German Residence Act, labor market participation is one key aspect of integration. Using the internet can increase migrants' wages or chances to find employment through more efficient employer-employee matching. For example, online job searching tools (e.g. Mang, 2012) or access to migrant online social networks (e.g. Edin, Fredriksson, and Åslund, 2003) are likely to improve match quality. Additionally, internet availability may encourage immigrants to take advantage of employment possibilities they would not have considered otherwise. The possibility of information retrieval through the internet significantly reduces transaction costs associated with residential relocation which is often necessary for finding employment (Komito and Bates, 2011). To capture this economic dimension of immigrant integration, I observe migrants' employment statuses, log gross monthly wages and whether they switched jobs between the pre and the post-broadband period. My results imply that migrants in households equipped with a broadband access are 5 percentage points more likely to be employed than those who cannot access the broadband internet. While this effect is only significant at the 10% level, in line with a causal interpretation, the respective coefficient is insignificant for inhabitants of municipalities located far away from the necessary broadband network infrastructure. None of the other economic integration variables are significantly related to broadband exposure.

Undoubtedly, another important dimension of integration are host country language

skills. Depending on whether immigrants use the internet to consume media and interact in their native language or in German, one can expect a negative or positive influence of broadband exposure on German language proficiency (e.g. Kissau, 2008). Surprisingly, neither self-reported German speaking nor writing proficiency are significantly related to broadband exposure in my analyses. Immigrants living in households with a broadband access, however, are more likely to regularly speak German at home than those without a broadband access. Among individuals in high-bandwidth locations, the respective probability is 8 percentage points higher and for all individuals together, irrespective of their residential location, it is 5 percentage points higher. As expected, for migrants in low-bandwidth locations, there is no significant difference in the probability to speak German at home between those individuals with and without a broadband access.

In order to measure migrants' participation in social everyday life, another goal stated in the Residence Act, I estimate regressions using four different dependent variables: an individual's frequency of performing voluntary activities, her probability to receive or visit Germans and, finally, her probability of having German friends. For example, Putnam (2000) argues that internet usage might substitute for real-life social interactions implying a negative effect on migrants' contacts with natives or on their general participation in the everyday life of the host country society. My analyses do not support this view. Although having a broadband access and using the broadband internet very frequently are related to a 9 percentage points and, respectively, 8 percentage points lower probability of having German friends, these effects are only significant in the baseline specification. In both, the high and the low bandwidth sample, the respective coefficients are insignificant. Furthermore, immigrants residing close to the necessary broadband network infrastructure increase their frequency of performing local voluntary activities by 16% and 20% of a standard deviation when having a broadband access at home or using the internet with a relatively high frequency, respectively. These effects are not measurable for individuals living in municipalities with bad broadband connectivity. The probabilities of receiving or visiting Germans in their houses are not related to broadband exposure.

Finally, I investigate those indicators also studied in the already existing literature on the relationship between internet usage and integration. Unlike the common finding that media consumption and communication with friends and family from origin countries affect ethnic identities, the willingness to remain in the host country, and life satisfaction, I do not find a causal link between these variables in my sample. Although there is a marginally significant and positive relationship in the baseline sample, the respective coefficients are insignificant for both samples used in my preferred specification.

Migrants' willingness to remain in Germany permanently and their life satisfaction are also not affected by broadband exposure. While life satisfaction, in line with the intuition that contact to family and friends alleviates stress from the migration experience, is positively correlated with broadband exposure in the baseline sample, this positive correlation remains even when investigating only those individuals living in low-bandwidth regions. Therefore it is likely that the respective baseline results are spurious.

Since broadband exposure might affect only subsamples of the immigrant population or depend on how exactly the internet is used, I also explore heterogeneity along different dimensions: first, the age of an individual could play an important role in how she is affected by broadband exposure. It is likely that younger migrants are more prone to use the internet in general or in a way which bears consequences for their integration (with respect to social contacts, compare e.g. Ye, 2005 and regarding language skills Bleakley and Chin, 2004). My results provide some evidence that it is primarily the young who use online job search tools. The positive effect of having a broadband access on employment probabilities decreases by one percentage point for each additional year of age. Age does not, however, significantly interact with broadband exposure effects for any other integration variable.

Second, using the internet for communication with friends and family or consumption of media from home countries, potentially even in the native language, is likely to have negative effects on German language proficiency, social contacts with Germans and on the degree of identification as German citizen. On the other hand, life satisfaction should be more positively affected by broadband exposure in Germany given that use pattern (Kissau, 2008). Since I do not observe the online activities of the individuals in my sample, I use OECD data on broadband penetration rates in origin countries as a proxy and interact them with the broadband exposure variables. Presumably, foreign-born individuals from countries with more households connected to the broadband network are more likely to use the internet for communication and media consumption in their native language. Indeed, for immigrants from countries with a 10 percentage points higher broadband penetration rate, the positive effect of a broadband access on the probability to usually speak German at home is 7 percentage points smaller. The interaction effect is similar in size for high-frequency internet users. However, both of the respective coefficients are only significant at the 10% level. Surprisingly, individuals in households equipped with a broadband access from countries with a 10 percentage points higher broadband penetration rate report an additional 30% of a standard deviation higher level of German speaking proficiency.

For the relationship between broadband exposure and social integration variables, the interaction effect is more consistent: although significant only in three models, it

is negative in all cases. Among immigrants living in households with a broadband access, those stemming from countries with a 10 percentage points higher broadband penetration rate are roughly 10 percentage points less likely to host Germans in their houses. Similarly, the probability to host and visit Germans in their houses is also 10 percentage points smaller for high-frequency internet users from countries with a 10 percentage points higher broadband penetration rate.

There is no significant interaction for the degree of feeling German, the wish to remain in Germany permanently or for life satisfaction.

The remainder of this paper is structured as follows. Section 2.2 describes the literature on the different dimensions of immigrant integration and explains how broadband exposure can affect each one of them. Section 2.3 gives an overview of Germany's migration history and its immigrant population. Section 2.4 describes the data analyzed throughout the paper, introduces the relevant variables used to capture integration and broadband exposure and shows some key characteristics of the immigrant sample. Section 2.5 introduces the baseline specification designed to assess the effects of broadband exposure on immigrant integration and presents the respective results for the different integration outcomes. Section 2.6 is intended to uncover some mechanisms behind the observed patterns of Section 2.5. It tests whether there is heterogeneity in the effects of broadband exposure on integration outcomes with respect to a migrant's age and to the broadband availability in her country of origin. In Section 2.7, I use the previously described instrumental variable to corroborate the causal interpretation of my baseline results. I perform a robustness check which does not rely on the instrumental variable in Section 2.8. Finally, Section 2.9 concludes.

2.2. Literature Review

The determinants of successful integration have been investigated extensively in various strands of literature. There are several dimensions frequently used to define whether integration can be considered successful.

First, economists have been primarily concerned with immigrants' labor market outcomes such as employment rates, wages (Chiswick, 1978; LaLonde, Topel, et al., 1992; Baker and Benjamin, 1994; Schoeni, 1997; Edin, Fredriksson, and Åslund, 2003; Antecol, Kuhn, and Trejo, 2006; Damm, 2009; Borjas, 2015), skill levels or occupational choices (Douglas, 1919; Green and Green, 1999; Abramitzky, Boustan, and Eriksson, 2014). Second, economists have also studied the determinants and effects of migrants' command of their host country's language (McManus, Gould, and Welch, 1983; Grenier, 1984; Kossoudji, 1988; Tainer, 1988; Rivera-Batiz, 1990; Chiswick, 1991; Rivera-

Batiz, 1992; Dustmann, 1994; Chiswick and Miller, 1995; Dustmann and Soest, 2001; Bleakley and Chin, 2004).

Third, research on the social aspect of integration investigates, for example, migrants' frequency of contacts with friends or acquaintances from the host country (Kissau, 2008; Hahn et al., 2019; Kosyakova and Brücker, 2020), memberships in local clubs and voluntary activity (Kissau, 2008) or the occurrence of intermarriages (Haug, 2004).

Fourth, to study whether and how ethnic identities of immigrants change after migration, researchers examine if individuals adopt new behaviors, values and attitudes, or, more generally, the extent to which migrants identify as citizen of their host country (Smither and Rodriguez-Giegling, 1982; Alba, 1990; Waters, 1990; Diehl and Schnell, 2006; Kissau, 2008).

Finally, another strand of literature views the well-being of immigrants or their willingness to remain in the host country permanently as indicators of successful integration and studies the respective determinants (Porter and Haslam, 2005; Berry and Hou, 2016; Akay, Bargain, and Zimmermann, 2017; Walther et al., 2020).

The potential effects of the internet on integration outcomes has already been studied for some of the previously mentioned dimensions. The majority of the respective research finds that using the internet to communicate with friends and family or to consume media from their respective home countries helps migrants to keep their heritage ethnic identity (Melkote and Liu, 2000; Panagakos, 2003; Hiller and Franz, 2004; Ye, 2005; Kissau, 2008; Komito and Bates, 2011; Kama and Malka, 2013; Son, 2015; Yin, 2015; Marat, 2016; Yoon, 2017). For example, Kissau (2008) states that this usage pattern has a negative effect on the extent to which migrants perceive themselves as being German or being part of the German society. In the same study, on the other hand, she finds that contacts to and media consumption from the origin country, enabled by the internet, positively affect life satisfaction and the willingness to remain in the host country, effects also mentioned by Ye (2005) and Kama and Malka (2013).

Regarding language skills, Kissau (2008) and Ye (2005) find that using the internet in the host country language correlates positively with host country language proficiency, however, as both authors also mention themselves, they cannot distinguish cause and effect of this relationship.

Finally, an interesting point is made by Komito and Bates (2011) who find that Polish and Filipino migrants in Ireland are more prone to internally migrate in their host countries to take advantage of employment possibilities in other cities because internet availability provides them with assurance to find all necessary information for the relocation process and social contacts online.

Although some of the previously mentioned studies suggest a connection between

internet exposure on the one hand and labor market integration and language proficiency on the other hand, generally, the literature on the relationship between the internet and integration lacks a thorough investigation of integration outcomes beyond ethnic identity and life satisfaction. This fact is surprising because the very same channels through which the internet affects ethnic identities and life satisfaction, that is media consumption and interpersonal communication, have been found to affect other outcomes relevant for integration.

First, studies on mass media exposure reveal significant impacts on a wide spectrum of individual characteristics such as educational attainment (Zavodny, 2006), occupational choices (Bjorvatn et al., 2020), political participation (Besley and Burgess, 2001; Strömberg, 2004; Kasper, Kogler, and Kirchler, 2015), violent behaviour (Anderson and Bushman, 2001), attitudes towards women (Jensen and Oster, 2009) or divorce probabilities (Chong and Ferrara, 2009). It is very plausible to think that mass media, being consumed over the internet, has the power to influence migrants' perception of the host country, its values and culture, thereby affecting ethnic identities and the willingness to participate in social and economic everyday life.

Second, similar to the previous intuition, peer effects in personal interactions, carried out online, could affect migrants' participation in the host country society (Funk, 2010; DellaVigna et al., 2016), their educational choices and therefore labor market outcomes (Bursztyn and Jensen, 2015; Chetty, Hendren, and Katz, 2016). Especially knowledge spillovers in immigrant networks, so-called ethnic enclaves, have been shown to disseminate a wide spectrum of information valuable to migrants (Kosyakova and Brücker, 2020). These range from particularities of the asylum process (Koser, 1997; Koser Akcapar, 2010) to those increasing job-worker match quality, thereby improving employment opportunities and raising wages (Edin, Fredriksson, and Åslund, 2003; Damm, 2009). Particular interesting in the context of this paper is the finding of Brekke and Brochmann (2015) who note that these kind of information are indeed spread via communication devices and social media.

Furthermore, a set of studies focusing on the effects of the internet on a general population bear implications for the integration success of immigrants in particular. In the context of labor market outcomes, Mang (2012) finds evidence that employer-employee match quality is increased through online job search possibilities. In line with this result, Kuhn and Mansour (2014) show that online job search significantly shortens unemployment spells. A more negative view is held by Putnam (2000) and Bauernschuster, Falck, and Woessmann (2014) who argue that internet usage, similar to TV consumption (Olken, 2009), might replace real-life social contacts thereby also decreasing the likelihood of civic engagement such as voluntary activity (Bauernschuster, Falck, and

Woessmann, 2014) or voting (Falck, Gold, and Heblich, 2014). This intuition, for migrants, suggests a negative effect of broadband exposure on contacts with Germans or active participation in German society.

With this paper, I augment the literature investigating the effects of internet exposure on immigrant integration in two ways: first, by studying a more comprehensive set of important integration indicators beyond ethnic identity and life satisfaction. Second, unlike previous research on the topic, I can base statistical inference on a more solid identification strategy and a significantly larger sample of foreign-born individuals.

2.3. Migration in Germany

In this section, I provide an overview of the migration movements to Germany since the second world war and the resulting composition of the immigrant population.

As of 2020, with roughly 18%, Germany had the 6th highest population share of foreign-born individuals among all EU countries. Table 2.3.1 provides an overview of the migrant population in Germany for the years 1973, 2000, 2008, 2014 and 2020. 1973 and 2020 represent the first and last year with available data, 2000 and 2008 roughly correspond to the pre and post-broadband era and 2014 marks the last year before the so-called refugee crisis.

Germany's high immigrant share is due to its history of institutional work migration. After the second world war, migration to Germany was limited to refugees and returnees who had been previously displaced. In the mid 1950s, however, economic growth led to a scarcity of workers which the German government sought to overcome through agreements with other countries that provided work migrants. These agreements led to considerable migration from countries such as Turkey, Italy, Greece and Yugoslavia from 1961 onward because the labor force could not be increased anymore with workers from East Germany due to the sealing off of the inner German border. Most of these so-called Guest Workers took blue-collar jobs in industrial production, involving physical labor, high health risks and low pay. Initially, it was planned that the Guest Workers would only stay in Germany for a certain period and then return to their home countries with their savings. Therefore, between the late 1950s and and the early 1970s, roughly 14 million foreigners migrated to Germany but at the same time 11 million also emigrated. Still, many work migrants prolonged their residence and subsequently also brought their families to Germany. Due to increased criticism of the Guest Worker program in the Germany society and the oil crisis of 1973, institutionalized migration came to an end. At this point, roughly 3.9 million immigrants (with foreign nationality), making up 6.3% of the population were living in Germany.

Table 2.3.1: Migrant population in Germany according to different definitions

Population *(in millions)*	**1973**	**2000**	**2008**	**2014**	**2020**
Germany	62.1	82.3	82.0	81.2	83.2
Foreign-born					15.0
(% of total)					*(18.1)*
With migration background[1]			14.6	16.3	20.0
(% of total)			*(18.0)*	*(20.1)*	*(24.0)*
Non-German	3.9	7.3	6.7	8.2	11.4
(% of total)	*(6.3)*	*(8.9)*	*(8.2)*	*(10.0)*	*(13.7)*
Share of nationalities *(% of non-German)*					
Turkish	*23.0*	*27.4*	*25.1*	*18.7*	*12.8*
Ex-Yugoslavian	*17.7*	*15.2*	*13.3*	*10.3*	*8.7*
Italian	*15.8*	*8.5*	*7.8*	*7.0*	*5.7*
Greek	*10.3*	*5.0*	*4.3*	*4.0*	*3.2*
Polish		*4.1*	*5.9*	*8.3*	*7.6*
Russian		*1.6*	*2.8*	*2.7*	*2.3*
Romanian		*1.2*	*1.4*	*4.4*	*7.0*
Afghan		*1.0*	*0.7*	*0.2*	*2.4*
Bulgarian		*0.5*	*0.8*	*2.2*	*4.0*
Syrian		*0.4*	*0.4*	*1.4*	*7.2*

Notes: This table shows the share of migrants, according to three different definitions, in the German population and the share of individuals with the ten most frequent nationalities in the non-German population as of 2000. **1:** Born without German nationality (non-German) or at least one parent is born without German nationality.
Data sources: Total population, population with migration background and non-German population data are taken from the Federal Statistical Office of Germany. The breakdown of population by nationalities for 1973 is taken from Höhne et al. (2014). Data on foreign-born individuals in Germany in 2020 are taken from Eurostat. Data for empty cells are not available.

In the early and mid 1980s, migration to Germany remained at low levels but surged in the beginning of the 1990s due to the fall of the Soviet Union, the wars in Yugoslavia and a crisis in the Kurdish inhabited part of Turkey. In 1992, roughly 1.5 million people (1.9% of the population) came to Germany but those numbers declined to 0.7 million arrivals in 2008 (0.9%). From then on, due to the ongoing European integration and the EU enlargements in 2004 and 2007, migration began to rise significantly with most immigrants stemming from Poland, Romania and Bulgaria. During the year 2014, roughly 1.5 million people (1.8%) had come to Germany.

In 2015, immigration to Germany peaked with 2.1 million (2.6%) arriving individuals because of the ongoing crisis in the Middle East and Northern Africa, especially the civil war in Syria. Until 2020, yearly migration rates slowly declined, to a yearly inflow of roughly 1.2 million people (1.4%), most of which came from Syria, Afghanistan, Irak, Iran and to a lesser extent from Sub-Saharan Africa, most notably Eritrea. As of 2020, almost a quarter of the German population had a migration background, more

than 18% were born abroad and roughly 14% did not have the German nationality. The largest group among non-Germans are still those with Turkish nationality, followed by Croatian, Polish, Syrian and Romanian citizens.[3]

2.4. Data

2.4.1. Variables and sources

The main dataset for my analysis is the German Socio Economic Panel (GSOEP or SOEP), an annual survey covering a representative share of German households. It is maintained by the German Institute for Economic Research, follows the same individuals since 1984 and contains a stable set of individuals who are surveyed annually. To mitigate panel attrition effects, new waves of households have been included into the sample at different points in time. The SOEP is intended to provide detailed information on the development of the German society in terms of, for example, income, wealth, labor market outcomes or life satisfaction. Especially convenient for the purposes of this paper, one of the explicit goals of the SOEP is to allow detailed studies of immigrants in Germany which is why this group of the population has been systematically oversampled for the first SOEP wave in 1984 (Krupp, 2008). All further analyses are based only on those individuals in the panel which were not born in Germany (irrespective of their nationality), that is those with migration experience. Note also that the minimum age to be included in the SOEP is 16 years.

To encompass the economic dimension of integration, I use SOEP variables on individuals' employment statuses, the log of their gross monthly wages[4] and whether they have switched jobs. I recode employment status which originally is measured on a categorical scale to account, for example, for part-time employment to a binary variable which equals unity if an individual is working in any type of job and zero else.

For measuring language proficiency, I rely on self-reported German speaking and writing skills, both measured on a 1 to 5 ordinal scale, with categories "Very good" (1), "Good" (2), "Fairly" (3), "Poorly" (4) and "Not at all" (5). I reverse the ranking and change the range to reach from 0 to 4 so that higher numbers indicate more successful integration and a value of zero corresponds to the category "Not at all".[5] Additionally, I include a categorical variable indicating which language migrants speak at home. The

[3]This section is based on information from the Federal Agency for Civic Education (*Bundeszentrale für Politische Bildung*) and on Feld et al. (2017).

[4]The gross monthly wage is calculated by dividing gross wages earned per year through the amount of months an individual was employed during that year.

[5]Table 2.4.2 shows that natives have higher values in integration variables than foreign-born individuals suggesting that higher values do indeed imply better integration.

answer options to the respective survey question are "only German", "part German, part origin country language" and "only origin country language". I transform this indicator to a binary variable such that the first two answer options are coded as one and the latter as zero. For all analyses involving language proficiency variables, I exclude foreign-born individuals from Switzerland or Austria as German is likely to be their native language and therefore not affected by broadband exposure.

I capture migrants' social contacts to Germans and their active participation in German society with a variable measuring the frequency of performing voluntary activities in local clubs and associations, two dummy variables indicating whether individuals received visits from or visited at least one German within the last year in their house, respectively, and finally a variable equalling unity if an individual has German friends and zero if not. The frequency of voluntary activity is measured on an ordinal 5 item scale with categories "Every day" (1), "Every week" (2), "Every month" (3), "Less frequently" (4) and "Never" (5). Similar to self-reported German proficiency, I reverse the ranking and change the range of the associated values.

For ethnic identity, I use a SOEP question which elicits foreign-born individuals' degree of feeling German on a 5 item ordinal scale with answer options "Completely" (1), "For the most part" (2), "In some respects" (3), "Barely" (4) and "Not at all" (5). Again, I reverse the ranking and change the reach from 0 to 4 such that higher numbers imply a higher degree of feeling German and a value of zero corresponds to the category "Not at all". To measure further integration dimensions, I use a dummy indicating an individual's wish to remain in Germany permanently and a variable on life satisfaction elicited on a scale from 0 to 10, with 10 representing the highest possible satisfaction. The selection of all variables is based on the relevant literature, especially on Kissau (2008). Table 2.A.1 in the appendix provides an overview of the various integration variables previously described, their original and modified scales and the SOEP waves in which they were collected.

Next to the integration variables, I obtain two indicators of broadband exposure from the SOEP. The first is measured at the household level and equals unity if an individual had a DSL access in her household in 2008. DSL was the main technology to access the broadband network in Germany and the variable was first collected in 2008. From now on, I will refer to this variable as broadband access. According to Falck, Gold, and Heblich (2014), residents of larger German cities were able to access the internet via a broadband connection for the first time in 1999. In mid 2002, the technology started to become available for a broader public. Therefore, broadband penetration was zero before 1999 and still close to zero before 2002. Consequently, I set broadband access to zero for all years before 2002 (Bauernschuster, Falck, and Woessmann, 2014 make the

same assumption). The second indicator varies at the individual level and elicits internet use frequency on an ordinal 5 item scale, with categories "Every day" (1), "Every week" (2), "Every month" (3), "Less frequently" (4) and "Never" (5). I perform several modifications so that the variable measures the frequency of using the broadband internet rather than the frequency of using the internet in general: first, again following Falck, Gold, and Heblich (2014), I set all values to 5, i.e. "Never", for observations before 2002. Second, I also set those values to 5 for individuals who live in households without a broadband access. Third, I set all values to missing for individuals for which information on the availability of a broadband access in the household is missing.[6] Finally, I recode the variable so that it equals unity if individuals use the internet "Every week" or "Every day" and zero otherwise. The previous step is taken to avoid using an ordinal regressor in my linear regressions. After these transformations, the variable categorizes individuals into high and low-frequency broadband users.

From a conceptual viewpoint, the frequency with which an individual uses the broadband internet is the most adequate broadband exposure indicator since having a broadband access only provides information on whether an individual *could* use the broadband internet and not how much she actually does use it. However, each of these measures have different advantages and caveats with respect to endogeneity (discussed in Section 2.5), such that using both should provide a comprehensive insight into the effects of broadband exposure on immigrant integration. Figure 2.A.1 in the appendix provides plots of the distribution of both broadband exposure variables as of 2008. The respective histogram for the broadband use frequency is based on all previously described transformations except of the last step making it a binary variable. Note also, since broadband access and use frequency are coded as zero in the pre-broadband period, changes in these variables will be similar to post-broadband levels.

On top of individual and household level measures of internet access, I include an instrumental variable in my dataset which predicts the internet bandwidth available to households across different municipalities in Germany. This variable was constructed by Falck, Gold, and Heblich (2014) and measures the distance of a municipality's geographic center to its so-called *Main Distribution Frame* (MDF), an essential part of the early German broadband infrastructure. Available internet bandwidth in a municipality declined with increasing distance from its MDF. Further explanations of the instrument, its relevance and conditional exogeneity are deferred to Section 2.7.

Additionally, I obtain the following control variables from the SOEP: the time an individual has spent in Germany, her country of origin, her municipality of residence,

[6]Out of the foreign-born population without a broadband access in 2008, only 23% percent report using the internet more often than "Never" and only 6% use it daily. 25% of those foreign-born individuals who never use the internet still have a broadband access.

her age, her gender, the number of years spent for education, and the number of children younger than 16 living in the same household. Given the information on time spent in Germany, I also construct a dummy variable assigning each individual to a specific cohort of migrants. The first cohort is made up of individuals who came to Germany before 1955, mostly those who had fled the war or ethnic Germans emigrating from the Soviet Union. The second and third cohort are the early and later Guest Workers, with the year 1961 distinguishing both groups. Immigrants who came to Germany between 1973 and 1989 are assigned to the fourth cohort. Those immigrating after the fall of the Soviet Union and during the war in ex-Yugoslavia between 1989 and 1995 make up the fifth cohort. Cohort six consists of individuals who immigrated between 1995 and 2003. The seventh cohort, spanning years 2004 to 2014 is intended to encompass immigrants from Eastern Europe after the EU's enlargements in 2004 and 2007. Since I have access only to the SOEP data until 2014, I cannot create a cohort representing immigrants from the Middle East and North Africa who came to Germany during the so-called Arab Spring. A detailed discussion of all control variables and their purpose follows in the next section.

From the website of the Federal Statistical Office of Germany, I obtain population and unemployment data on the municipality level as further control variables.[7] I compute the female population share, the share of individuals aged between 18 and 65 and the share of individuals older than 65. The data are available for the year 2001 and from 2008 onward. For 2001, observations are missing for the federal states of Baden-Württemberg and Saarland. I also use a 17 item variable categorizing each German municipality according to its degree of agglomeration provided by Falck, Gold, and Heblich (2014) who in turn obtained it from the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2007). The categorization is based on a municipality's population density, its area used for settlement and a centrality index calculated from the proximity to populous regions and employment possibilities. Municipalities of type 1 to 8 are considered agglomerated, those of types 9 to 17 as less agglomerated.

Finally, I also collect national level data on the share of households connected to the broadband network for all OECD countries for years 2005 to 2008 from the OECD's *Broadband Portal*. The broadband penetration rate is used to test whether broadband exposure effects on migrants' integration outcomes in Germany are affected by broadband

[7]The Federal Statistical Office does not provide municipality level data on unemployment rates but only on the number of unemployed (and the Federal Employment Agency publishes unemployment rates only on the county level). I approximate the unemployment rate by dividing the number of unemployed though the number of individuals aged 18 to 65. The official definition would require using the labor force as divisor.

availability in their origin countries. As with German broadband exposure variables, I assume that the broadband penetration rates in origin countries were zero before 2002.

Not all of the previously discussed variables, especially those obtained from the SOEP, are available for the same years (for the availability of integration indicators, compare Table 2.A.1). To simplify my empirical analysis and increase statistical accuracy (compare Bauernschuster, Falck, and Woessmann, 2014 and Falck, Gold, and Heblich, 2014), I collapse my panel data such that only two periods remain: averages across years 1999, 2000 and 2001 represent the pre-broadband era whereas averages across years 2006, 2007 and 2008 make up the post-broadband period. Only for the variable indicating whether an individual plans to remain in Germany permanently, I have to use 2010 values for the post-broadband period. Summary statistics are provided in the following subsection.

2.4.2. Immigrant sample

The baseline sample which I will analyze throughout this paper is comprised of all foreign-born individuals who where surveyed in the SOEP in both the pre and the post-broadband period and for whom information on the availability of a broadband access is non-missing in 2008. I do not impose an age restriction and the youngest individual in the sample is 16 years old in 1999. Among the foreign-born individuals surveyed in both periods, some data are missing for the integration variables. This is because, on the one hand, wages were only elicited from employed survey participants (and not coded as zero) and, on the other hand, some outcomes were only asked from subsamples of all foreign-born SOEP participants. Therefore, regression samples vary across models with different dependent variables. To provide consistent summary statistics and since broadband availability will serve as regressor for all integration indicators, I restrict the baseline sample to those individuals for whom this variable is non-missing.[8] The number of foreign-born individuals surveyed in the pre-broadband period is 3687. 2184 of these were also surveyed in the post-broadband period. Finally, 1548 individuals remain for whom information on broadband availability is non-missing. The most important immigrant groups in the sample, in terms of origin countries (not nationalities), are individuals born in Turkey, Ex-Yugoslavia, Poland, Italy and Russia.

Table 2.4.1 shows the share and composition of foreign-born individuals, those with migration background and non-Germans in the SOEP sample. Since only individuals surveyed in both periods are included in the sample, the share of foreign-born and individuals with migration background does not change between periods. Note that,

[8]There are 6 individuals for whom information on broadband use frequency is missing although information on broadband access is available.

Table 2.4.1: Migrants in the SOEP according to different definitions

Sample	**Pre-broadband** *(1999-2001)*		**Post-broadband** *(2006-2008)*
Total SOEP[1]		13,668	
Foreign-born		1,548	
(% of total)		*(11.3)*	
With migration background[2]		2,556	
(% of total)		*(18.7)*	
Non-German	968		889
(% of total)	*(7.4)*		*(6.5)*
Share of nationalities /origin countries	*(% of foreign-born)*	*(% of non-German)*	*(% of non-German)*
Turkish	*21.5*	*33.1*	*31.1*
Ex-Yugoslavian	*12.0*	*17.6*	*18.1*
Polish	*9.7*	*1.2*	*1.0*
Italian	*8.5*	*16.2*	*18.0*
Russian	*7.9*	*1.5*	*1.0*
Kazakhstan	*7.2*	*0.2*	*0.1*
Romanian	*5.1*	*0.3*	*0.2*
Greek	*4.4*	*8.5*	*9.6*
Spanish	*2.0*	*3.5*	*3.6*
Austrian	*1.8*	*2.5*	*3.2*

Notes: This table shows the share of migrants, according to three different definitions, in the SOEP sample and the share of individuals with the ten most frequent (in the pre-broadband period) origin countries in the foreign-born and non-German sample, respectively. **1:** The total SOEP sample is limited to those individuals who were surveyed in both, the pre and the post-broadband period and for whom information on broadband availability in 2008 is non-missing. **2:** The definition of the SOEP is different than that from the Federal Statistical Office of Germany: individuals with own migration experience or with at least one parent with migration experience.

although the SOEP is intended to be a representative sample of the German population, neither immigrant shares nor the percentages of nationalities and origin countries match those of the overall German population as reported in Table 2.3.1 for the year 2000. In the pre-broadband period, with 7.4%, the share of non-Germans among survey respondents is 1.5 percentage points lower than in the population. This number reduces to 6.5% in 2008 which is 1.7 percentage points less than the respective share in the overall population.[9]

Table 2.4.2 additionally provides an overview of some key differences between the foreign-born and native sample in the SOEP for the pre-broadband period. Only the

[9]Two reasons explain this observation: first, my sample consists of three SOEP waves, ranging from 1999 to 2001. Second, when restricting the sample, attrition and missing data distort the proportion and composition of the non-German sample relative to its population counterpart. However, this should pose no problem since my goal is to investigate the relationship between broadband exposure and integration and not to make statements about the migrant population in general. Also, the ranking of nationalities by shares in the SOEP is very similar to that of the overall population.

Table 2.4.2: Comparison between native and foreign-born individuals

Variable	Mean		Δ Foreign-born - Native	
	Native[1]	Foreign-born	Unconditional	Age-gender-education adjusted
Employment status	0.60	0.57	-0.03**	-0.06***
			(0.01)	(0.01)
Gross monthly wage	3279.02	2907.84	-371.18***	-8.57
			(62.18)	(48.67)
Frequency of voluntary activity	0.59	0.24	-0.35***	-0.30***
			(0.02)	(0.02)
Life satisfaction	7.17	7.11	-0.06*	0.04
			(0.03)	(0.04)
Age	45.92	43.71	-2.20***	
			(0.37)	
Male	0.48	0.48	0.00	
			(0.01)	
Nr. of children in h.h.	0.76	1.15	0.39***	0.29***
			(0.02)	(0.02)
Years of education	11.93	10.57	-1.36***	
			(0.06)	
Population *(in 1000)*	137.67	165.26	27.59***	46.73***
			(6.96)	(7.23)
Unemployment rate in municipality of residence *(in %)*	6.94	5.72	-1.22***	-1.14***
			(0.08)	(0.09)
Degree of agglomeration in municipality of residence[2]	0.48	0.58	0.11***	0.14***
			(0.01)	(0.01)
Broadband access in h.h.	0.54	0.49	-0.05***	-0.03**
			(0.01)	(0.01)
Broadband use frequency	0.41	0.30	-0.11***	-0.07***
			(0.01)	(0.01)
Distance of municipality of residence to MDF *(in km)*	1.75	1.41	-0.34***	-0.41***
			(0.03)	(0.03)
Broadband penetration in country of origin	0.18	0.09	-0.09***	-0.09***
			(0.00)	(0.00)

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table compares characteristics across the native and the foreign-born population in the SOEP. Columns 2 and 3 show the means of the respective characteristics for each group. Columns 4 and 5 show coefficients from regressing the characteristics on a dummy indicating if an individual is foreign-born. Column 5 is based on regressions including the age of an individual (as linear, squared and cubic term), the gender and the years of education as control variables. **1:** The native sample is comprised of those individuals who were surveyed in the pre and post-broadband period, for whom information on broadband access is non-missing and who are non-foreign-born with German-born parents. **2:** Consistent with the definition of the BBSR (2007), agglomerated municipalities are those of type 1 to 8.

broadband related variables refer to post-broadband values. For the purpose of this comparison only, I exclude those individuals from the non-foreign-born baseline sample who were born in Germany but have parents who are foreign-born.[10] Since some

[10]Out of the 12,120 non-foreign-born individuals in the SOEP who were surveyed in the pre and post-

integration variables (e.g. language proficiency) were not elicited from individuals satisfying this definition of natives, they are not included in Table 2.4.2. For the selected variables, columns two and three show the means for each group, respectively and column four lists the respective differences. Column five also shows mean differences, however adjusted for age, gender and education to control for potential differences not stemming from an individual's migration history but from systematic differences in the composition of the immigrant sample and the native population. Each cell shows the coefficient of a dummy variable indicating whether an individual was foreign-born (as opposed to native, according to the above definition) from regressions of the different variables on that dummy and the age of an individual (as a cubic), the gender and her years of education.

The patterns are mostly as expected: the foreign-born population is slightly but significantly younger, by about 2 years, and less educated, i.e. about 1.4 years less years of schooling, than the natives. Also, immigrants tend to live in households with more young children, on average 0.4. Gender shares, however, do not significantly differ between both groups. On average, migrants live in cities with about 28,000 more inhabitants which are 11 percentage points more likely to be agglomerated according to the BBSR (2007) definition and exhibit a 1.2 percentage point lower unemployment rate. For those integration variables available for both samples, the foreign-born score significantly worse than the natives: They are 3 percentage points less likely to be employed and, on average, earn 370 Euros less of gross monthly salary. Their frequency of performing voluntary activities is also 40% of a standard deviation lower. Still, foreign-born individuals report an average life satisfaction which is only 4% of a standard deviation lower than that of natives. This difference is also only marginally significant at the 10% level. In terms of broadband exposure, migrants are 5 percentage points less likely to live in a household equipped with a broadband access and 11 percentage points less likely to be high-frequency broadband users. On top of that, the average broadband penetration rate in their home countries is 9 percentage points lower than that in Germany. Foreign-born individuals, however, live in municipalities located, on average, 340 meters closer to their MDFs than those municipalities inhabited by natives. This observation is consistent with the fact that, generally, agglomerated municipalities are located more closely to their MDFs (compare Section 2.7).

When controlling for age, gender and education, most of these differences become smaller in absolute terms but still remain significant. Only life satisfaction and wages are similar among both groups after including control variables. Notably, conditional mean

broadband period and for whom data on broadband access is non-missing, 1008 have foreign-born parents (i.e. they have a migration background according to the SOEP definition) and are therefore excluded from the non-foreign-born sample.

differences are larger in absolute terms for employment probabilities, the population, degree of agglomeration and the MDF distance of the residence municipality. While migrants are significantly less likely to be employed, those that have a job earn comparable wages as their German colleagues. Also, consistent with research on the topic, foreign-born individuals seem to self-select into more densely populated, agglomerated regions offering employment possibilities.

Table 2.4.3 provides summary statistics for the changes of the integration and broadband exposure indicators and all previously mentioned control variables for individuals of the baseline sample. Tables 2.A.2 and 2.A.3 in the appendix contain the pre and post-broadband values of the respective variables. On average, most of the integration outcomes have worsened between periods, though only some changes are significant: notably, the share of employed individuals and their wages have decreased between periods. This dynamic might be related to the consequences of the financial crisis of 2007. Some evidence for this interpretation is given by the fact that the average unemployment rate of the sample municipalities increased significantly by 1.4 percentage points. In line with this observation, life satisfaction also declined significantly. The frequency of voluntary activity, however, significantly increased over time.

2.5. Baseline Analysis

2.5.1. Model specification

In this subsection, I will present the baseline specification to test the effect of broadband exposure on migrants' integration outcomes. Each of the following subsections will be devoted to one of the integration dimensions and the respective indicators discussed in Section 2.4.1.

The exact form of the specification is a modification of the wage assimilation equation presented by Borjas (2015).[11] Additional control variables are chosen based on Bauernschuster, Falck, and Woessmann (2014):

$$\Delta Int_i = \beta_0 + \beta_1 \Delta BB_i + \boldsymbol{\Delta X_i' \beta_2} + \boldsymbol{C_i' \beta_3} + \epsilon_i \tag{2.1}$$

ΔInt_i is the change in the integration outcome of migrant i between the post and the pre-broadband period. The constant β_0 is equivalent to a time fixed-effect which absorbs any general trends in outcome variables between periods such as, for example,

[11] Alternatively, to account for the binary and ordinal scales of the integration measures, I could also use (ordered) logit or probit models. However, these would not allow me to control for time-invariant factors which are very likely to bias the obtained results. Therefore, I do not opt for this alternative.

Table 2.4.3: Summary statistics, changes

Variable **Δ Post - pre-broadband**	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
Integration indicators					
Employment status	-0.03	0.41	-1.00	1.00	1548
Log gross monthly wage	-0.31	0.59	-2.90	2.22	761
Changed job	0.18	0.38	0.00	1.00	1548
Usual language is German	0.01	0.31	-1.00	1.00	1025
Spoken German proficiency	0.06	0.61	-2.50	2.50	1015
Written German proficiency	0.08	0.70	-2.50	3.00	1014
Frequency of voluntary activity	0.06	0.64	-3.00	3.50	1526
Received Germans	-0.00	0.40	-1.00	1.00	805
Visited Germans	-0.03	0.42	-1.00	1.00	806
Has German friends	-0.03	0.56	-1.00	1.00	1420
Degree of feeling German	0.10	1.08	-3.00	4.00	820
Wish to remain in Germany	-0.01	0.34	-1.00	1.00	1162
Life satisfaction	-0.43	1.39	-5.67	6.67	1547
Broadband access	0.47	0.50	0.00	1.00	1548
Broadband use frequency	0.28	0.45	0.00	1.00	1542
Broadband penetration in country of origin	0.10	0.07	0.04	0.31	841
Distance to MDF *(in km)*	1.42	1.20	0.08	9.33	1489
Control variables					
Years of education	0.11	0.66	-2.00	5.00	1466
Nr. of children in h.h.	-0.12	0.97	-8.33	4.17	1545
Population *(in 1000)*	1.50	102.77	-1223.20	1287.34	1481
Female population share *(in %)*	-0.08	0.54	-4.35	5.49	1481
Share of working age population *(in %)*	-2.49	2.61	-13.69	4.56	1481
Share population older than 65 *(in %)*	1.99	1.42	-6.06	13.06	1481
Unemployment rate *(in %)*	1.40	1.40	-5.71	6.52	1481
Cohort					
1950-1954	0.01	0.11	0.00	1.00	1461
1955-1961	0.03	0.18	0.00	1.00	1461
1962-1973	0.29	0.45	0.00	1.00	1461
1974-1898	0.35	0.48	0.00	1.00	1461
1990-1995	0.25	0.43	0.00	1.00	1461
1996-2003	0.07	0.25	0.00	1.00	1461
Sample size			1548		

Notes: This table shows summary statistics for changes of all integration, broadband exposure and control variables used throughout the paper except of the squares and cubics of an immigrant's age and her time of residence in Germany. The pre and post-broadband levels of the last two variables can be found in Tables 2.A.2 and 2.A.3, respectively. The job change, the distance to MDF variable and the cohort fixed-effects are not based on changes but levels. Changes are calculated by subtracting pre-broadband from post-broadband values. The sample consists of all foreign-born individuals who were surveyed at least once in the pre and post-broadband period and for whom information on broadband availability is non-missing. Summary statistics for the country of origin fixed-effect can be found in Table 2.4.1.

the overall increase in unemployment. BB_i is either the indicator variable for the availability of broadband access or for being a frequent broadband user. Depending on the exact integration outcome, each of the broadband indicators might introduce a differ-

ent type of endogeneity to the respective regression, and the empirical model is chosen to address as many of them as possible. For example, using first-differences with a constant makes inference robust to linear age-effects, like younger migrants using the broadband internet more frequently and finding it easier to learn the German language. Other potential individual fixed-effects that influence broadband exposure and integration outcomes alike, e.g. the socio-economic status of parents, are also taken care of by exploiting the panel structure of my data.

To address sources of endogeneity originating from time-variant factors, I include $\boldsymbol{\Delta X_i}$, a vector of controls which contains changes in the following variables: an immigrant's age and the time she lives in Germany (both squared and cubed), the number of years of education, the number of children under 16 living in the household, the population, unemployment rate, female population share, share of working-age population and share of population older than 65 in the municipality of residence.

On top of their linear effects, age, and the number of years since arrival in Germany both might affect broadband exposure and integration non-linearly. For example, young individuals who are not yet part of the labor market and the old who already retired have more time to spare for private social contacts and lower or no wages as compared to middle-aged working individuals. At the same time, the very young and old might be more prone to use the broadband internet, again, due to more availability of free time. Years of education have undeniable effects on labor market outcomes and language proficiency but also could be correlated to broadband exposure. It might be more likely for well-educated individuals to use a (then) new technology such as broadband internet. Having children increases the probability to get into contact with other, potentially German, parents and at the same time is likely to increase the probability of having a broadband access at home, for example because children pressure their parents to buy one. The municipality level control variables are included to proxy for local socio-economic conditions potentially affecting the integration possibilities of the immigrant population.

$\boldsymbol{C_i}$ is a vector of different fixed-effects. It includes the arrival cohort, a country of origin and two regional fixed-effects: one indicating federal states and the other the municipality type according to its degree of agglomeration. The regional fixed-effects are based on the last observed residence in the pre-broadband period (rather than the last observed residence in the post-broadband period) to avoid potential endogeneity between broadband access and relocation decisions.[12] Changes in the municipality level

[12]Recall that the location of a municipality determines the internet bandwidth available to its households, the very fact Falck, Gold, and Heblich (2014) use to construct their instrumental variable. I follow Bauernschuster, Falck, and Woessmann (2014) who also base regional fixed-effects on pre-broadband period residence.

control variables, however, are based on actual residences. The set of fixed-effects is included because, as Borjas (2015) suggests, different migration dates, countries of origin and residential locations might affect the *rate* of immigrant integration. Instead (or on top) of the two regional fixed-effects, I could also use a county fixed-effect. However, individuals in my sample live in 259 distinct counties and including these as a fixed-effect would absorb too much variation and bear the risk of overfitting the model.

Finally, standard errors are clustered at the pre-broadband residence municipality level. Despite using first-differences and including all of the previously mentioned control variables, reverse causality might still bias the results of specification (2.1). For example, individuals who find employment between periods will have less time to use the internet resulting in a spurious, negative correlation between changes in labor market participation and broadband use frequency. On the other hand, migrants learning the German language faster or those making many German-speaking friends might find it easier to sign a contract with an internet provider. I address these concerns in Section 2.7.

2.5.2. Economic integration

In this section, I present results of regression equation (2.1) for the economic integration indicators. Recall that broadband exposure facilitates online job search which in turn increases employer-employee match quality (Mang, 2012), shortens unemployment spells (Kuhn and Mansour, 2014) and, through knowledge spillovers, reduces transaction costs associated to job search (Komito and Bates, 2011). Especially ethnic networks, likely also accessible via the internet (Brekke and Brochmann, 2015), have been shown to disseminate job information, further improving employer-employee match quality. These considerations imply that broadband exposure should affect employment probabilities and log wages positively.

Table 2.5.1 shows results of estimating equation (2.1) using the three economic integration indicators as dependent variables. Coefficients are mostly insignificant. Only for those individuals who obtained a broadband access, the probability of being employed is 5 percentage points higher than for those without it. However, the effect is just marginally significant at the 10% level. The fact that the probability of switching jobs is not affected by broadband exposure suggests that the internet might help migrants to find employment but is less useful for transitioning between different jobs. In line with this intuition, wages are also not related to broadband exposure: if the internet were to increase employer-employee match quality through online search, the respective coefficients should be significant and positive. Recall that, since wages were only elicited from employed individuals, regressions cannot pick up any positive effect on wages

Table 2.5.1: Baseline results, economic integration

Dependent variable:	Δ Employment status		Δ Log wage		Changed job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.05*		0.05		0.02	
	(0.02)		(0.05)		(0.02)	
Δ Broadband use frequency		0.03		0.06		0.01
		(0.03)		(0.05)		(0.03)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	1279	1274	656	653	1279	1274
R-sq.	0.14	0.14	0.24	0.24	0.13	0.13

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of economic integration indicators as dependent variables. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

stemming from individuals who find new jobs after having been unemployed.

Note further that the most pressing endogeneity concern, given the first-difference specification and controls, is likely a reverse causality problem implying that individuals without employment or with lower wages might have more time to use the internet. In regressions (1) to (4), this intuition would be reflected as negative coefficients, so that the obtained coefficients might be downward biased.

2.5.3. Language proficiency

The following section presents results of the baseline specification using indicators of German language proficiency as dependent variables. A priori, it is not obvious whether broadband exposure should have a positive or negative effect on German language skills or use. Migrants using the internet to consume German media or interact with Germans will likely experience a positive influence whereas exposure and communication in the native language is likely to be detrimental for German language proficiency (Ye, 2005; Kissau, 2008). The direction of effects could bear some insights into how immigrants use the broadband internet.

Table 2.5.2 presents the respective baseline results. Individuals with a broadband access experienced a 5 percentage points greater increase in the likelihood to speak Ger-

Table 2.5.2: Baseline results, language proficiency

Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.05**		-0.01		0.08	
	(0.03)		(0.05)		(0.05)	
Δ Broadband use frequency		-0.01		-0.05		0.01
		(0.02)		(0.06)		(0.06)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	865	862	857	854	857	854
R-sq.	0.08	0.07	0.15	0.16	0.11	0.11

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of German proficiency indicators as dependent variables. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

man at home. Intuitively, being exposed to German-language broadband internet might familiarize household members with the German language and, additionally, normalize its everyday use. In turn, the insignificance of the broadband use frequency variable might be based on the fact that, especially for larger households, the household-wide used language lies outside the scope of a single individual and is likely determined by the person with the lowest German proficiency.

Generally, since results for the other models are insignificant, evidence for an effect of broadband exposure on German language proficiency is limited.

2.5.4. Social integration

This section presents results for regressions using different indicators of immigrants' contacts to Germans as dependent variables. Research suggests that broadband exposure could affect interactions with Germans in two ways. First, similar to the intuition of better employer-employee match quality, a fast internet access or frequent use might enable immigrants to find new friends and acquaintances online (Ye, 2005). However, second, virtual contacts, or time spent online in general, might substitute for face-to-face interactions such that broadband exposure reduces migrants' contact with Germans

Table 2.5.3: Baseline results, social integration

Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Regressors:								
Δ Broadband access	0.07**		0.06		0.01		-0.09**	
	(0.04)		(0.04)		(0.04)		(0.04)	
Δ Broadband use frequency		0.08**		0.01		0.02		-0.08**
		(0.04)		(0.04)		(0.04)		(0.04)
Fixed-effects								
Cohort	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	1262	1257	677	673	678	674	1178	1173
R-sq.	0.12	0.12	0.19	0.18	0.18	0.18	0.12	0.12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of social integration indicators as dependent variables. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

(Putnam, 2000; Bauernschuster, Falck, and Woessmann, 2014).

At least the baseline results presented in Table 2.5.3 cannot resolve which of both intuitions applies. On the one hand, foreign-born individuals with a broadband access in their house and high-frequency broadband users exhibit an 11% and 13% of a standard deviation higher frequency of performing voluntary activities than their counterparts, respectively. On the other hand, these two groups are 9 percentage points and 8 percentage points less likely to have German friends, respectively.

In Sections 2.6 and 2.7, I will investigate whether these seemingly counterintuitive results might be based on heterogeneous effects of broadband exposure or, potentially, endogeneity biases.

2.5.5. Ethnic identity and other integration measures

The literature on internet and integration suggests that migrants predominantly use the internet to stay in touch with their origin country such as to maintain and strengthen their ethnic identity (compare, amongst many others, Melkote and Liu, 2000 or Marat, 2016). Kissau (2008) even reports a negative impact of internet exposure on migrants' identification as part of the German society. On the other hand, contact to the home

Table 2.5.4: Baseline results, other integration indicators

Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.12		0.03		0.17*	
	(0.10)		(0.03)		(0.10)	
Δ Broadband use frequency		0.20*		-0.01		0.21**
		(0.11)		(0.03)		(0.10)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	695	693	989	985	1279	1274
R-sq.	0.15	0.15	0.14	0.14	0.09	0.09

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of miscellaneous integration indicators as dependent variables. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

country might affect more general indicators of integration such as the willingness to stay in Germany indefinitely or overall life satisfaction positively (Kissau, 2008; Kama and Malka, 2013). Furthermore, the literature on peer effects and mass media exposure implies that broadband use in German might familiarize migrants with the way of life in Germany and thus have integrative effects (see, for example, Chong and Ferrara, 2009 or Jensen and Oster, 2009).

Table 2.5.4 indicates that, if anything, identification as German citizen seems to be positively affected by broadband exposure. High-frequency broadband users report a roughly 19% of a standard deviation higher degree of feeling German than foreign-born individuals who use the broadband internet infrequently. Note, however, that this effect is only significant at the 10% level. In line with the literature, individuals from households with broadband availability and high-frequency users score 13% to 15% of a standard deviation higher on the scale of overall life satisfaction, respectively.

2.6. Heterogeneous Effects

In this section, I explore whether there is heterogeneity in the effects of broadband exposure with respect to two dimensions. First, I conjecture that the age of an immigrant is a

significant factor which influences effects since, generally, the young are probably more inclined to use new media (Statistisches Bundesamt, 2015). On top of that, younger migrants might use the internet in a systematically different way than the old, therefore also creating scope for heterogeneous effects. In particular, it seems likely that broadband exposure has significantly more positive effects for young individuals regarding economic integration because first, the young look for jobs more frequently in general and second, they are more likely to look for them online. Furthermore, age might significantly interact with broadband exposure effects on German language use and proficiency because younger individuals find it easier to learn new languages (Bleakley and Chin, 2004) and are more likely to practice their German skills through online communication, for example, via online chats or games (Ye, 2005). For the same reasons, age might interact with the effect of broadband exposure on immigrants' social integration.

The second variable which potentially alters the effects of broadband exposure in Germany is the broadband availability in migrants' origin countries. Intuitively, better broadband availability in the home country will allow immigrants to use the internet in their native language and to communicate with friends and family from their home countries. The relevant literature suggests that broadband exposure might have significantly more negative effects on language proficiency, contacts to Germans and the degree of identification as German for those migrants who predominantly use the internet in their native language. On the other hand, better opportunities to keep in touch with family and friends from the origin country likely imply more positive effects of broadband exposure on life satisfaction (compare, for example, Kissau, 2008).

2.6.1. Age

I begin the analysis by investigating whether the age of an immigrant interacts with the effect of broadband exposure on integration outcomes. I estimate regression equation (2.1) and add an interaction term between the demeaned age and the respective broadband exposure variable. I also include a main effect for age. Coefficients on the main broadband regressor represent marginal effects for an immigrant of average pre-broadband period age of 44 years.

Economic integration The results shown in the upper panel of Table 2.6.1 indicate that especially younger migrants profit from broadband exposure, possibly because they are more likely to search for jobs online. Obtaining broadband access has a smaller impact on the employment probability of old workers than on the employment probability of young workers. The same is true for high-frequency broadband users although the coefficient on the respective interaction term is only significant at the 10% level. Each

Table 2.6.1: Heterogeneous effects, age

Economic integration						
Dependent variable:	Δ Employment status		Δ Log wage		Changed job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.03		0.02		0.01	
	(0.02)		(0.05)		(0.02)	
Δ Broadband access × Age	-0.01***		-0.01		-0.00	
	(0.00)		(0.00)		(0.00)	
Δ Broadband use frequency		0.01		0.03		0.01
		(0.03)		(0.06)		(0.03)
Δ Broadband use frequency × Age		-0.00*		-0.00		0.00
		(0.00)		(0.01)		(0.00)
Fixed-effects	✓	✓	✓	✓	✓	✓
Obs.	1279	1274	656	653	1279	1274
R-sq.	0.14	0.14	0.24	0.24	0.13	0.13
Language proficiency						
Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:						
Δ Broadband access	0.06*		0.01		0.09	
	(0.03)		(0.06)		(0.06)	
Δ Broadband access × Age	0.00		0.01		0.00	
	(0.00)		(0.00)		(0.00)	
Δ Broadband use frequency		-0.01		-0.01		0.05
		(0.03)		(0.07)		(0.08)
Δ Broadband use frequency × Age		-0.00		0.00		0.00
		(0.00)		(0.01)		(0.01)
Fixed-effects	✓	✓	✓	✓	✓	✓
Obs.	865	862	857	854	857	854
R-sq.	0.08	0.07	0.16	0.16	0.12	0.11
Social integration								
Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Regressors:								
Δ Broadband access	0.06*		0.06		0.01		-0.09**	
	(0.04)		(0.04)		(0.05)		(0.04)	
Δ Broadband access × Age	-0.00		0.00		-0.00		-0.00	
	(0.00)		(0.00)		(0.00)		(0.00)	
Δ Broadband use frequency		0.05		0.04		0.01		-0.09**
		(0.05)		(0.04)		(0.04)		(0.04)
Δ Broadband use frequency × Age		-0.00		0.00		-0.00		-0.00
		(0.00)		(0.00)		(0.00)		(0.00)
Fixed-effects	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	1262	1257	677	673	678	674	1178	1173
R-sq.	0.12	0.12	0.19	0.19	0.19	0.19	0.12	0.12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of integration indicators as dependent variables. The broadband exposure variables are interacted with the demeaned age of a migrant. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

additional year above the average age decreases the effect of having a broadband access or being a high-frequency user on the probability of being employed by 0.5 and 0.4 percentage points, respectively.

For log gross monthly wages and the probability of switching jobs, however, there are no significant heterogeneous effects with respect to age. One potential explanation for this finding might be that the young are more prone to use the internet for job search, however they are not necessarily more likely to use the internet to switch to a higher paying employment.

Language proficiency Younger individuals learn languages faster and are more likely to use the internet in ways beneficial for language skill acquisition. Both intuitions suggest that their language skills might be significantly more positive affected by broadband exposure. Surprisingly, however, as the results in the middle panel of Table 2.6.1 indicate, this is not the case in my sample.

Social Integration Differences in online activity patterns of the young such as chatting and gaming could be beneficial for contacts with Germans. It is not clear, however, whether those contacts are complements or substitutes of face-to-face interactions. In my analysis I cannot find any evidence in support of either intuition: the coefficients on the interaction terms shown in the lower panel of Table 2.6.1 are all insignificant.

2.6.2. Broadband penetration in origin country

In this subsection, I will investigate whether broadband availability in an immigrant's country of origin influences broadband exposure effects on integration in Germany. I interact each broadband exposure indicator with the demeaned change in the broadband penetration of a migrant's heritage country. I also add the respective main effect. One might be worried that broadband penetration rates of origin countries pick up other effects such as, for example, cultural proximity to Germany. This effect, however, is controlled for by the first-difference specification and, in case one is additionally worried about linear trends, by the origin country fixed-effect. Table 2.B.1 in the appendix shows broadband penetration rates for the most frequent origin countries in my sample. Since they are based on OECD resources, data for most Ex-Yugoslavian countries (except of Slovenia), Russia, Kazakhstan and Romania are missing. The broadband penetration rates vary from 4% in Turkey to 17% in Austria.

Language proficiency I begin the analysis with language proficiency indicators. The upper panel of Table 2.6.2 shows that broadband exposure has (more) negative effects

Table 2.6.2: Heterogeneous effects, broadband penetration in country of origin

Language proficiency						
Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.04		-0.02		0.09	
	(0.03)		(0.06)		(0.06)	
Δ Broadband access	-0.71*		1.76**		1.28	
× Δ Broadbandp. in or. count.	(0.42)		(0.86)		(1.00)	
Δ Broadband use frequency		-0.01		-0.05		-0.02
		(0.03)		(0.06)		(0.07)
Δ Broadband use frequency		-0.68*		0.63		0.28
× Δ Broadbandp. in or. count.		(0.36)		(0.83)		(1.05)
Fixed-effects	✓	✓	✓	✓	✓	✓
Obs.	587	584	578	575	579	576
R-sq.	0.09	0.09	0.17	0.16	0.11	0.10

Social integration								
Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Regressors:								
Δ Broadband access	0.13**		0.04		-0.01		-0.06	
	(0.05)		(0.05)		(0.05)		(0.06)	
Δ Broadband access	0.09		-1.03*		-0.97		-0.66	
× Δ Broadbandp. in or. count.	(0.86)		(0.61)		(0.60)		(0.69)	
Δ Broadband use frequency		0.12**		0.02		0.02		-0.08
		(0.05)		(0.04)		(0.04)		(0.06)
Δ Broadband use frequency		0.58		-0.94*		-1.24**		-0.65
× Δ Broadbandp. in or. count.		(0.93)		(0.48)		(0.50)		(0.78)
Fixed-effects	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	720	716	517	513	518	514	681	677
R-sq.	0.13	0.13	0.19	0.19	0.18	0.18	0.09	0.09

Other integration indicators						
Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(15)	(16)	(17)	(18)	(19)	(20)
Regressors:						
Δ Broadband access	0.20		0.03		0.18	
	(0.13)		(0.04)		(0.13)	
Δ Broadband access	3.18		-0.17		0.15	
× Δ Broadbandp. in or. count.	(1.97)		(0.61)		(1.85)	
Δ Broadband use frequency		0.30**		-0.01		0.13
		(0.13)		(0.04)		(0.13)
Δ Broadband use frequency		2.59		0.60		-0.28
× Δ Broadbandp. in or. count.		(1.64)		(0.56)		(1.73)
Fixed-effects	✓	✓	✓	✓	✓	✓
Obs.	447	445	618	614	728	724
R-sq.	0.19	0.19	0.14	0.14	0.10	0.09

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of integration indicators as dependent variables. The broadband exposure variables are interacted with the demeaned broadband penetration in the migrant's country of origin. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

on the usual language spoken at home for individuals from countries with higher broadband availability. A 10 percentage point increase in the broadband penetration rate of the origin country, decreases the effect of both broadband exposure indicators in Germany by roughly 7 percentage points. Note, however, that the coefficients are significant only at the 10% level. Surprisingly, the effect of broadband availability on German speaking proficiency is significantly stronger for migrants from countries with good broadband internet. A 10 percentage point increase in the origin country broadband penetration rate above the average is related to an increase in the effect of broadband access by roughly 30% of a standard deviation.

Social integration The results shown in the middle panel Table 2.6.2 indicate that indeed, better broadband availability in origin countries might deter migrants from forming social ties with Germans. A 10 percentage point increase in the broadband penetration rate of the origin country decreases the effect of both broadband exposure indicators on the probabilities to host Germans in their homes by roughly 10 percentage points. Coefficients are significant at the 10% level only. Additionally, the effect of being a high-frequency broadband user on the probability to visit Germans in their houses deceases by 12 percentage points for each 10 percentage point increase in the broadband penetration rate of the origin country. While the former two coefficients are only marginally significant, their magnitudes still suggest that broadband effects on the probabilities of hosting and visiting Germans are negative for individuals from countries with an average broadband penetration rate.

Ethnic identity and other integration measures Although the literature suggests that broadband exposure in the native language should bear consequences for ethnic identities and have positive effects on the willingness to remain in the host country and on life satisfaction, I cannot replicate these results in my analysis. None of the interaction terms in the lower panel of Table 2.6.2 have significant coefficients.

2.7. Instrumental Variable Approach

Despite the first-difference specification and the included control variables, my results may still be subject to omitted variable biases and reverse causality. There might be trends simultaneously correlated with changes in broadband use frequency or the decision to acquire a broadband access on the one hand and the integration success on the other. For example, changes in some integration indicators, such as contacts to Germans, are likely to improve language proficiency and also raise the probability to

buy a broadband access. Furthermore, reverse causality might bias coefficients because changes in integration could affect changes in broadband use frequency, for example when individuals get a job.

Therefore, I use an instrumental variable constructed by Falck, Gold, and Heblich (2014) which proxies for *changes* in the German regional broadband penetration rate. To understand the relevance of the instrument, consider the following paragraph which is based on their paper: when broadband was introduced in Germany, it relied on the existing telephone infrastructure. A bottleneck for the available bandwidth at the household level was the distance of a household to its *Main Distribution Frame* (MDF), a unit which was located in a dedicated building and intended to link the national telephone network with local street cabinets. The greater the distance, the lower the bandwidth available to a household. In particular, if a household was located further away to its MDF than 4.2km, available bandwidth dropped below broadband internet speed. To construct their instrumental variable, Falck, Gold, and Heblich (2014) collected data on the distance of all German municipalities (where the location of a municipality is its geographic center) to their MDFs. This variable is likely to be exogenous because for telephone service quality the distance of an MDF to its municipality did not matter. Instead, other considerations, in particular available lots, had a bigger impact on location choice. According to the authors, the biggest concern regarding the exclusion restriction is that MDFs were more likely to be located closer to agglomerated areas. If living in these regions systematically affects integration outcomes, the instrument would be endogenous within the context of this paper. The municipality level variables of the baseline specification are chosen to control for this possibility. For further explanations on the instrument see Falck, Gold, and Heblich (2014) or Campante, Durante, and Sobbrio (2018) who use a similar approach.

Unfortunately, the distance of the municipality of residence to its MDF is neither suitable for instrumenting the changes in the presence of a broadband access in the household nor for individual changes in broadband use frequency. The respective F-statistics are well below the rule of thumb value of 10 across all specifications. This is most likely due to the fact that individuals are not aware of their municipality's distance to its MDF and decide to subscribe for broadband services irrespective of this consideration. Tables 2.C.1 to 2.C.4 in the appendix show results from two-stage least squares (2SLS) regressions where I instrument both broadband exposure variables with the distance of an individual's municipality of residence to its MDF. Equations (2.2) and (2.3) describe the respective first and second stage specifications. Across all integration outcomes, none of the IV coefficients are significant. The respective Kleibergen-Paap rk Wald F statistics take values well below two and all coefficients on the instrumental

variable in the first stage are close to zero and insignificant.

I can still use the instrumental variable in two different ways. First, I estimate reduced form versions of the baseline specification, replacing the potentially endogeneous broadband exposure measures with the distance of an individual's municipality of residence to its MDF. Equation (2.4) in the appendix describes the respective specification and Tables 2.C.5 to 2.C.8 show the results for all integration indicators. Each table also includes three modifications of the instrument, all of which are binary dummies grouping individuals according to the distances of their residence municipalities to their respective MDFs. Since a distance of 4.2km marks the threshold at which bandwidth is too low for accessing the internet with broadband speed, one of the dummy variables equals unity for individuals living in municipalities further away than 4.2km from their MDFs and zero else. However, only 4% of immigrants in my sample live in regions so far away from their MDFs. To compromise between the problems caused by small sample biases and ensuring to create a sample of individuals with poor broadband connectivity, I also show results for dummies splitting the sample at the median and the 75% percentile of the distance variable. For the baseline sample, the median distance of a municipality to its MDF is 1095 meters and the 75% percentile corresponds to 1829 meters. Recall, that these values might vary due to missing data in regression samples. Except for changes in German writing proficiency, migrants' degrees of feeling German and life satisfaction, coefficients of the reduced form specifications are insignificant. Individuals in regions further away from MDFs than 50% or 75% of the municipalities in the sample, exhibit marginally significant, positive changes in their German writing proficiency. Migrants in cities further away from MDFs than the 4.2km threshold value report significantly higher degrees of feeling German and values of life satisfaction. Only for the latter outcome, the continuous distance variable also yields a significant positive coefficient. Since distance is a negative predictor of regional broadband availability, the latter two result stand in stark contrast to the positive coefficients of the respective baseline specifications.

Given the weak predictive power of the instrument, I opt for implementing a second, placebo type approach which is my preferred specification: migrants living in regions with low bandwidth availability will not be able to use the broadband internet no matter if they have a broadband access or not. If the significant effects of the presence of a broadband access and broadband use frequency on integration are driven by biases or reverse causality, coefficients should also be significant for individuals living in municipalities far away from their MDFs. If, however, it is indeed broadband exposure driving the results, coefficients should lose significance for individuals in those regions. To test whether significant coefficients of the baseline specification in Section 2.5 are unbiased,

I split the respective regression samples into individuals living in regions close to their MDF and into individuals living in regions far away from their MDF. Conditional exogeneity of the instrument implies that this sample split should only affect regression results through its effect on actual broadband availability. For this approach, I divide the sample at the 75% percentile of the distance variable. Tables 2.C.9 to 2.C.12 in the appendix show results for the same analysis, using the median distance as threshold value. I do not show results for individuals living in regions further away from their MDFs than 4.2km because sample sizes in this case vary between 13 and 36 individuals which is too low for accurate inference. The following subsections describe the results of the placebo approach for each integration dimension separately.

2.7.1. Economic integration

Coefficients of the baseline specification implied that a broadband access in the household was related to a 5 percentage points increase in employment probabilities. None of the other economic integration indicators was significantly affected by any of the broadband exposure measures. Table 2.7.1 repeats the baseline specification on both of the previously described samples. For individuals living in municipalities with good broadband connectivity, employment probabilities are also 5 percentage points higher when they have broadband access. Furthermore, unlike in the baseline specification, migrants with a broadband access and high-frequency broadband users earn about 11 log points higher wages. Although the coefficients are only significant at the 10% level, the fact that their counterparts in the sample of individuals with poor broadband connectivity are smaller and insignificant, supports a causal interpretation of broadband exposure effects.

Using the medium distance to a municipality's MDF yields qualitatively similar results. Table 2.C.9 shows that individuals with a broadband access in regions with high broadband penetration rates have a 8 percentage points higher employment probability than individuals in similar regions but without a broadband access. Here, the coefficient is significant even at the 5% level. For migrants living far away from their MDFs, the estimation produces a much smaller and insignificant coefficient. This pattern, however, is not observable for log gross monthly wages.

Table 2.7.1: Placebo estimation results, economic integration

Distance to MDF:	$\leq 75\%$ percentile						$> 75\%$ percentile					
Dependent variable:	Δ Employment status		Δ Log wage		Change job		Δ Employment status		Δ Log wage		Change job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	0.05*		0.11*		0.02		0.03		0.01		-0.01	
	(0.03)		(0.06)		(0.03)		(0.05)		(0.08)		(0.07)	
Δ Broadband use frequency		0.02		0.11*		-0.00		0.06		0.02		0.04
		(0.03)		(0.06)		(0.03)		(0.05)		(0.08)		(0.08)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	965	962	491	490	965	962	302	300	156	154	302	300

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of economic integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 75% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

2.7.2. Language proficiency

The results presented in Table 2.5.2 suggested that immigrants' probability to speak German at home is, albeit only marginally significant, 5 percentage points higher when they have a broadband access. Table 2.7.2 shows that this relationship is even stronger for individuals living in municipalities close to their MDFs. With 8 percentage points, the coefficient is larger than in the baseline specification and furthermore significant at the 1% instead of the 10% level. Additionally, having a broadband access at home increases German writing proficiency by roughly 16% of a standard deviation although the respective coefficient is only marginally significant. In line with a causal interpretation of these effects, the same coefficients are much smaller (even negative) and insignificant in the sample of individuals with poor broadband connectivity.

Table 2.C.10 in the appendix supports these results, at least regarding the probability of speaking German at home: using the median distance as threshold, the coefficient on broadband access in the sample of individuals living close to their MDFs is slightly smaller and only significant at the 10% level. Still, for migrants with poor broadband connectivity, there is no significant relationship between the presence of a broadband access and the probability to usually speak German. Regarding German writing proficiency, as in the baseline specification, none of the models yield significant results.

2.7.3. Social integration

With respect to social integration, the baseline analysis revealed inconclusive outcomes. While broadband availability and high-frequency use were related to roughly 12% of a standard deviation increase in the frequency of performing voluntary activities, they also implied an 8 percentage points lower probability of having German friends (see Table 2.5.3). Table 2.7.3 confirms the positive effect of broadband exposure on the frequency of performing voluntary activities: for individuals in the sample with good broadband availability, broadband access and high-frequency use are related to 16% and 19% of a standard deviation higher frequencies of performing such activities, respectively. Both coefficients are significant at the 5% level and become negative and insignificant in the sample of regions far away from their MDFs.

Table 2.7.2: Placebo estimation results, language proficiency

Distance to MDF:	$\leq 75\%$ percentile						$> 75\%$ percentile					
Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency		Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	0.08***		0.01		0.11*		-0.06		-0.08		0.02	
	(0.03)		(0.06)		(0.06)		(0.07)		(0.12)		(0.14)	
Δ Broadband use frequency		0.01		-0.03		0.05		-0.07		-0.08		-0.00
		(0.03)		(0.07)		(0.08)		(0.06)		(0.11)		(0.13)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	649	647	641	639	641	639	209	208	208	207	208	207

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of German proficiency indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 75% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.7.3: Placebo estimation results, social integration

Distance to MDF:	$\leq 75\%$ percentile								$> 75\%$ percentile							
Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends		Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Regressors:																
Δ Broadband access	0.10**		0.08*		0.03		-0.07		0.08		-0.03		-0.06		-0.09	
	(0.04)		(0.04)		(0.05)		(0.05)		(0.08)		(0.09)		(0.12)		(0.10)	
Δ Broadband use frequency		0.12**		0.03		0.01		-0.06		0.07		-0.09		0.04		-0.11
		(0.05)		(0.05)		(0.05)		(0.05)		(0.09)		(0.06)		(0.08)		(0.10)
Fixed-effects																
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	950	947	505	502	506	503	887	884	300	298	161	160	161	160	282	280

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of social integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 75% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Results for the probability of having German friends, however, remain inconclusive. The coefficients on broadband exposure indicators are insignificant in both samples, potentially because of too small sample size. Next to these outcomes, the probability of a foreign-born individual to host a German in her house is 8 percentage points higher for migrants with a broadband access, living in areas with high broadband penetration rates. The respective coefficient is only marginally significant but, in line with a causal effect, loses its significance completely when estimated on the sample of individuals with poor broadband connectivity.

Table 2.C.11 in the appendix produces slightly different results. Using the median distance value to split samples, the frequency of performing voluntary activities is only significantly higher for high-frequency broadband users among individuals living in municipalities closer to their MDFs. The same coefficient is insignificant for migrants living in areas with poor broadband connectivity. Outcomes for the probability of having German friends seem to suggest that baseline results are biased: the respective coefficients on the broadband exposure indicators are insignificant for individuals in cities with good broadband connectivity but become significant for those in areas with low broadband penetration rates.

2.7.4. Ethnic identity and other integration indicators

According to the baseline model, degree of identification as German and life satisfaction are positively influenced by broadband exposure. However, a migrant's degree of feeling German was only related to the frequency of broadband use and the respective coefficient was just marginally significant (see Table 2.5.4). Among individuals in areas with good broadband connectivity, there does not seem to be any relationship between these integration outcomes and broadband exposure. The respective coefficients in Table 2.7.4 are all insignificant. On top of that, for individuals who live in areas with low broadband penetration rates, broadband use frequency is significantly related to the willingness to remain in Germany forever and to life satisfaction.

When splitting the sample at the median distance (see Table 2.C.12), those individuals living in areas far away from their MDFs and having a broadband access or being high-frequency broadband users, report significantly higher levels of feeling German. Different from the results shown in Table 2.7.4, broadband access seems to be positively related to life satisfaction for individuals in areas with good broadband connectivity but not for those who live in cities with poor connectivity.

Generally, however, given the inconsistent results across my specifications, it seems to be the case that broadband exposure, unlike stated in the literature, does not significantly influence ethnic identities or life satisfaction.

Table 2.7.4: Placebo estimation results, other integration indicators

Distance to MDF:	$\leq 75\%$ percentile						$> 75\%$ percentile					
Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction		Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	0.09		0.04		0.15		0.29		0.06		0.24	
	(0.13)		(0.03)		(0.12)		(0.17)		(0.08)		(0.17)	
Δ Broadband use frequency		0.16		0.04		0.17		0.26		-0.11*		0.33*
		(0.13)		(0.03)		(0.12)		(0.18)		(0.05)		(0.17)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	520	518	742	738	965	962	165	165	239	239	302	300

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of miscellaneous integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 75% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

2.8. Robustness check

In this section, I perform a robustness check to corroborate the findings of the placebo estimations from the previous section.

If unobserved trends or reverse causality affect estimations only in the sample of individuals living in regions with good broadband connectivity, the conditional exogeneity assumption of the instrument would be violated. For example, this situation could occur if, first, migrants who learn German are more likely to buy a broadband access but, second, only do so in areas where broadband is technically available. In that case, regressions for individuals in regions far away from their MDFs would yield insignificant results without implying that results are unbiased for those migrants living in areas with good broadband connectivity. To rule out this possibility, I employ a second placebo strategy which does not rely on the instrumental variable. Instead of using spatial variation of broadband penetration, I use time variation: I repeat the baseline specifications on the same sample of individuals, however, I investigate changes in the respective variables before the broadband period. In particular, I use changes between the pre-broadband period and the period comprised by the years 1992 and 1993. Still, the broadband exposure variables will equal unity for those individuals who bought a broadband access or became high-frequency broadband users between the pre and the post-broadband period. However, since broadband was not available or close to zero between 1992 and 2001, the respective coefficients should be insignificant. If they are not, broadband exposure indicators are endogenous, implying that baseline results might be inconsistent.

Tables 2.D.1 to 2.D.4 in the appendix show results for all integration indicators. Apart from using changes between different periods, the underlying regressions differ only from those of the baseline specification in that the share of the working age population, the share of individuals older than 65 and the unemployment rate at the municipality level are not included as controls because the respective data are not available. Among those specifications which yield significant coefficients in the baseline regressions, almost none yield significant results in the pre-broadband periods. Exceptions are the models using the probability of speaking German in the household and of having German friends as dependent variables. Obtaining a broadband access between the pre and the post-broadband period is related negatively to changes in the probability of speaking German at home between both pre-broadband periods. This result strengthens the finding of the respective baseline regression that broadband availability increases migrants' probability to usually speak German at home. Similarly, the coefficient on the broadband access regressor in the model predicting the probability of having German friends also has the opposite sign as in the respective baseline model. Still, evidence

from Section 2.7 suggests that the relationship between broadband exposure and the probability of having German friends is likely prone to biases.

Some coefficients in Tables 2.D.1 to 2.D.4 are significant although their counterparts in the baseline specifications are not. This is likely due to the fact that for regressions of changes in the pre-broadband periods, I cannot use the full set of control variables.

2.9. Conclusion

In this paper, I extend the literature on the relationship between internet exposure and immigrant integration by studying a wide spectrum of relevant integration outcomes. I show that broadband exposure can help migrants integrate into the German society, at least with respect to some aspects of integration. Compared to immigrants who cannot access the broadband internet, foreign-born individuals living in households equipped with a broadband internet access are 5 percentage points more likely to be employed, have a 5 percentage points to 8 percentage points higher probability of usually speaking German at home and engage in local voluntary activities 16% of a standard deviation more frequently.

I corroborate these results using an instrumental variable which predicts available internet bandwidth at the municipality level. Although, due to too low F-statistics, estimation of a two-stage least squares model is not feasible, I use the instrument for a placebo-type approach: in line with a causal interpretation of my results, broadband exposure affects integration outcomes only in regions with high bandwidths and not in municipalities with poor broadband connectivity.

The positive effect of broadband exposure on employment probabilities is likely attributable to the fact that young migrants access the internet more frequently in general and are more prone to use online job search tools. Interacting broadband exposure regressors with an individual's age shows that effects on employment probabilities decrease by one percentage point for each year of age.

As suspected in the related literature, exposure to the German language on the internet, normalizes and therefore increases German language use at home. To proxy the likelihood with which migrants use the internet in their native or the German language, I obtain data on the number of households connected to the broadband network in the origin countries of the foreign-born individuals in my sample. A 10 percentage point increase in the origin country broadband penetration rate reduces the positive effect of German broadband exposure on the probability to speak German at home by roughly seven percentage points.

For some integration outcomes, I cannot reproduce the findings or conjectures of

the respective literature. While some research on social capital suggests that broadband exposure might have negative effects on social contacts and active participation in society, my results indicate a positive influence on local voluntary activity and no impact on migrants' number of German friends. There is, however, some evidence that those individuals using the internet predominantly to stay in contact with their countries of origin, are less likely to visit or receive Germans in their houses. Furthermore, I can reject the hypothesis, postulated in qualitative research on the topic, that broadband exposure affects migrants ethnic identities or life satisfaction. Results of my placebo estimations indicate no significant causal link between these variables.

Appendix

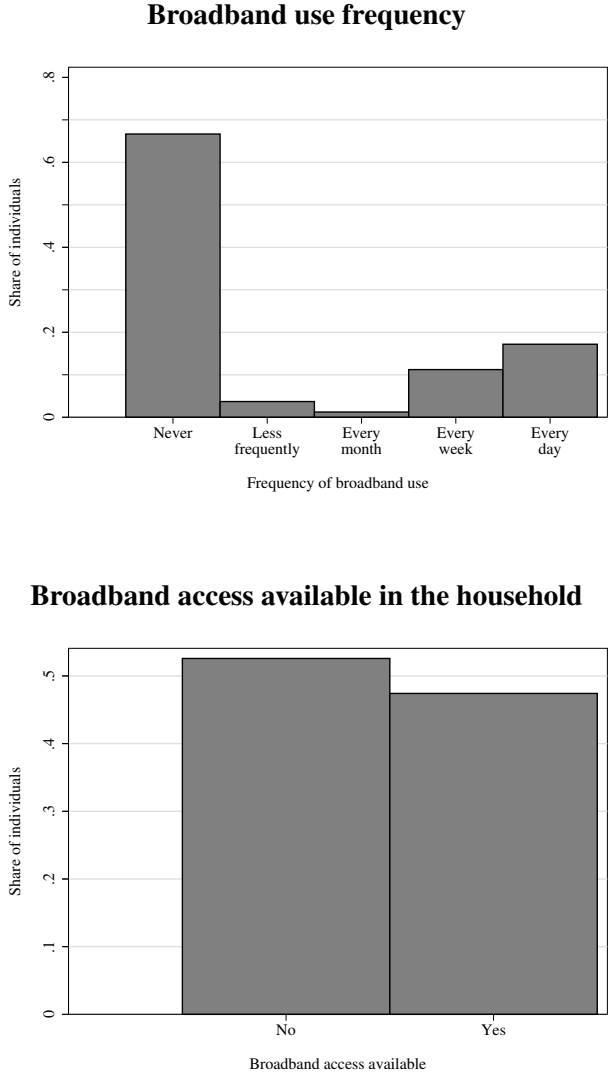
2.A. Appendix to Section 2.4

Table 2.A.1: Dimensions and indicators of integration

Dimension and indicator	**Original scale** *[min-max]*	**Modified scale** *[min-max]*	**Available years**
Economic			
Employment status	Categorical	Binary	'99-'08
Gross monthly wage	Cardinal	Cardinal (log)	'99-'08
Changed job	Categorical	Binary	'99-'08
Language proficiency			
Language usually spoken at home	Categorical	Binary	'99-'01, '07, '08
Spoken German proficiency	Ordinal *[5-1]*	Ordinal *[0-4]*	'99, '01, '07, '08
Written German proficiency	Ordinal *[5-1]*	Ordinal *[0-4]*	'99, '01, '07, '08
Social			
Frequency of voluntary activity in clubs, associations or community service	Ordinal *[5-1]*	*Ordinal [0-4]*	'99, '01, '07, '08
Received visits from Germans in own home	Binary		'99, '01, '07
Visited Germans in their home	Binary		'99, '01, '07
Has German friends	Categorical	Binary	'01, '06
Ethnic identity and other			
Degree of feeling German	Ordinal *[5-1]*	Ordinal *[0-4]*	'99, '01, '10
Wish to remain in Germany permanently	Binary		'99-'01, '06-'08
Life satisfaction	Ordinal *[0-10]*		'99-'08

Notes: This table shows SOEP indicators for different dimensions of integration. The classification of indicators and names of classes/dimensions are based on Kissau (2008).

Figure 2.A.1: Empirical distribution of broadband exposure measures



Notes: These figures show the empirical distribution of both broadband exposure measures in the post-broadband period.

Table 2.A.2: Summary statistics, pre-broadband levels

Variable Pre-broadband	Mean	Std. Dev.	Min.	Max.	Obs.
Integration indicators					
Employment status	0.58	0.45	0.00	1.00	1548
Log gross monthly wage	7.70	0.76	4.70	9.74	996
Changed job	0.00	0.00	0.00	0.00	1548
Usual language is German	0.80	0.34	0.00	1.00	1253
Spoken German proficiency	2.79	0.96	0.00	4.00	1232
Written German proficiency	2.28	1.22	0.00	4.00	1230
Frequency of voluntary activity	0.23	0.60	0.00	3.00	1528
Received Germans	0.83	0.33	0.00	1.00	830
Visited Germans	0.80	0.35	0.00	1.00	831
Has German friends	0.60	0.49	0.00	1.00	1471
Degree of feeling German	2.18	1.26	0.00	4.00	1227
Wish to remain in Germany	0.77	0.37	0.00	1.00	1253
Life satisfaction	7.15	1.46	0.67	10.00	1547
Broadband access	0.00	0.00	0.00	0.00	1548
Broadband use frequency	0.00	0.00	0.00	0.00	1548
Broadband penetration in country of origin	0.00	0.00	0.00	0.00	845
Distance to MDF *(in km)*	1.42	1.20	0.08	9.33	1489
Control variables					
Age	43.58	14.58	17.00	84.00	1548
Nr. of children in h.h.	1.13	1.25	0.00	10.33	1545
Years of education	10.67	2.41	7.00	18.00	1510
Years since migration	19.32	11.35	0.00	50.00	1461
Population *(in 1000)*	159.65	283.56	0.41	1726.36	1483
Female population share *(in %)*	51.26	0.96	46.28	56.17	1483
Share of working age population *(in %)*	67.08	2.15	59.59	74.26	1483
Share population older than 65 *(in %)*	17.19	2.20	9.43	28.20	1483
Unemployment rate *(in %)*	5.72	2.25	1.34	21.82	1483
Sample size			1548		

Notes: This table shows summary statistics for the pre-broadband levels of all variables used throughout the paper and an individual's age as well as her length of residence in Germany (years since migration). The sample consists of all foreign-born individuals who were surveyed at least once in the pre and post-broadband period and for whom information on broadband availability is non-missing.

Table 2.A.3: Summary statistics, post-broadband

Variable **Post-broadband**	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
Integration indicators					
Employment status	0.54	0.46	0.00	1.00	1548
Log gross monthly wage	7.30	0.81	4.38	9.49	873
Changed job	0.18	0.38	0.00	1.00	1548
Usual language is German	0.79	0.38	0.00	1.00	1146
Spoken German proficiency	2.81	0.96	0.00	4.00	1150
Written German proficiency	2.31	1.23	0.00	4.00	1151
Frequency of voluntary activity	0.29	0.69	0.00	3.50	1546
Received Germans	0.88	0.32	0.00	1.00	1500
Visited Germans	0.83	0.37	0.00	1.00	1503
Has German friends	0.57	0.50	0.00	1.00	1487
Degree of feeling German	2.26	1.29	0.00	4.00	954
Wish to remain in Germany	0.76	0.38	0.00	1.00	1334
Life satisfaction	6.72	1.56	0.33	10.00	1548
Broadband access	0.47	0.50	0.00	1.00	1548
Broadband use frequency	0.28	0.45	0.00	1.00	1542
Broadband penetration in country of origin	0.10	0.07	0.04	0.31	841
Distance to MDF *(in km)*	1.41	1.16	0.04	9.33	1496
Control variables					
Age	50.43	14.62	23.00	91.00	1548
Nr. of children in h.h.	1.01	1.20	0.00	10.33	1548
Years of education	10.83	2.50	7.00	18.00	1477
Years since migration	26.16	11.43	6.00	57.00	1461
Population *(in 1000)*	161.30	284.10	0.15	1739.23	1496
Female population share *(in %)*	51.18	0.92	46.38	54.99	1496
Share of working age population *(in %)*	64.60	3.13	50.79	73.18	1496
Share population older than 65 *(in %)*	19.19	2.31	10.37	31.78	1496
Unemployment rate *(in %)*	7.13	2.71	1.60	25.38	1496
Sample size			1548		

Notes: This table shows summary statistics for the post-broadband levels of all variables used throughout the paper and an individual's age as well as her length of residence in Germany (years since migration). The sample consists of all foreign-born individuals who were surveyed at least once in the pre and post-broadband period and for whom information on broadband availability is non-missing.

2.B. Appendix to Section 2.6

Table 2.B.1: Summary statistics, broadband penetration in country of origin

Broadbandp. in country of origin **Δ Post - pre-broadband**	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Overall	0.09	0.06	0.01	0.31
Turkey	0.04	0.01	0.02	0.07
Ex-Yugoslavia[1]	0.16	0.04	0.10	0.19
Poland	0.06	0.01	0.02	0.08
Italy	0.14	0.02	0.11	0.16
Russia				
Kazakhstan				
Romania				
Greece	0.05	0.02	0.01	0.06
Spain	0.14	0.02	0.10	0.15
Austria	0.17	0.01	0.13	0.19
Observations		1225		

Notes: This table shows summary statistics for the broadband penetration in an immigrant's country of origin. Data on non-OECD countries are missing. **1:** Based on Slovenia only (five individuals).

2.C. Appendix to Section 2.7

2.C.1. Two-stage least squares regressions

Specification

$$\Delta BB_i = \alpha_0 + \alpha_1 Dist_i + \boldsymbol{\Delta X_i'\alpha_2} + \boldsymbol{C_i'\alpha_3} + u_i \tag{2.2}$$

$$\Delta Int_i = \beta_0 + \beta_1 \widehat{\Delta BB_i} + \boldsymbol{\Delta X_i'\beta_2} + \boldsymbol{C_i'\beta_3} + \epsilon_i \tag{2.3}$$

Equation (2.2) describes the first stage and (2.3) the second stage. $Dist_i$ is the distance of the municipality of residence of individual i to its MDF and $\widehat{\Delta BB_i}$ is the predicted value of either of the broadband exposure measures.

Results

Table 2.C.1: Two-stage least squares estimation results, economic integration

Dependent variable:	Δ Employment status		Δ Log wage		Changed job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Endogenous regressors:						
Δ Broadband access	-0.40		0.83		-0.58	
	(0.89)		(1.23)		(1.02)	
Δ Broadband use frequency		-0.30		0.65		-0.42
		(0.60)		(0.85)		(0.62)
First stage						
Coefficient on instrument	0.00	0.00	0.00	0.00	0.00	0.00
Kleibergen-Paap rk Wald F statistic	0.60	1.38	0.73	1.49	0.60	1.38
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	1279	1274	656	653	1279	1274

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of 2SLS regressions (equations (2.2) and (2.3)) using changes between the post and pre-broadband values of economic integration indicators as dependent variables. The main regressors are instrumented by the distance of an individual's municipality of residence to its MDF. Controls include changes between post and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.2: Two-stage least squares estimation results, language proficiency

Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Endogenous regressors:						
Δ Broadband access	2.33		4.85		8.48	
	(37.77)		(175.69)		(106.83)	
Δ Broadband use frequency		-0.22		0.11		0.77
		(0.70)		(1.16)		(1.50)
First stage						
Coefficient on instrument	-0.00	0.00	0.00	0.00	0.00	0.00
Kleibergen-Paap rk Wald F statistic	0.00	0.81	0.00	0.94	0.01	1.08
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	865	862	857	854	857	854

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of 2SLS regressions (equations (2.2) and (2.3)) using changes between the post and pre-broadband values of German proficiency indicators as dependent variables. The main regressors are instrumented by the distance of an individual's municipality of residence to its MDF. Controls include changes between post and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.3: Two-stage least squares estimation results, social integration

Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Endogenous regressors:								
Δ Broadband access	0.12		-0.08		-0.04		-0.25	
	(1.19)		(0.91)		(1.02)		(1.29)	
Δ Broadband use frequency		0.08		-0.04		-0.01		-0.14
		(0.84)		(0.56)		(0.63)		(1.00)
First stage								
Coefficient on instrument	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kleibergen-Paap rk Wald F statistic	0.57	1.43	0.49	1.92	0.51	1.97	0.79	1.61
Fixed-effects								
Cohort	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	1262	1257	677	673	678	674	1178	1173

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of 2SLS regressions (equations (2.2) and (2.3)) using changes between the post and pre-broadband values of social integration indicators as dependent variables. The main regressors are instrumented by the distance of an individual's municipality of residence to its MDF. Controls include changes between post and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.4: Two-stage least squares estimation results, other integration indicators

Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Endogenous regressors:						
Δ Broadband access	-2.44		-0.65		7.55	
	(6.84)		(1.75)		(9.83)	
Δ Broadband use frequency		4.75		-0.32		5.59
		(15.28)		(0.58)		(4.72)
First stage						
Coefficient on instrument	-0.00	0.00	0.00	0.00	0.00	0.00
Kleibergen-Paap rk Wald F statistic	0.20	0.09	0.26	1.55	0.60	1.38
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	695	693	989	985	1279	1274

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of 2SLS (equations (2.2) and (2.3)) regressions using changes between the post and pre-broadband values of miscellaneous integration indicators as dependent variables. The main regressors are instrumented by the distance of an individual's municipality of residence to its MDF. Controls include changes between post and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

2.C.2. Reduced form regressions

Specification

$$\Delta Int_i = \beta_0 + \beta_1 Dist_i + \boldsymbol{\Delta X_i' \beta_2} + \boldsymbol{C_i' \beta_3} + \epsilon_i \tag{2.4}$$

Table 2.C.5: Reduced form results, economic integration

Dependent variable:	Δ Employment status				Δ Log wage				Changed job			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Distance	-0.00				0.02				-0.01			
	(0.01)				(0.02)				(0.01)			
Distance>Median		0.00				0.03				-0.01		
		(0.02)				(0.04)				(0.02)		
Distance>75% percentile			-0.03				0.04				-0.00	
			(0.02)				(0.05)				(0.02)	
Distance>4.2km				0.02				0.03				-0.04
				(0.04)				(0.11)				(0.04)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	1630	1630	1630	1630	741	741	741	741	1630	1630	1630	1630
R-sq.	0.12	0.12	0.12	0.12	0.25	0.25	0.25	0.25	0.12	0.12	0.12	0.12

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.4) using changes between the post and pre-broadband values of economic integration indicators as dependent variables. For each dependent variable, the first model is estimated using the distance to a municipality's MDF as continuous variable. The second and third models use a dummy indicating whether a migrant lives in a region further away from its MDF than 50% or 75% of individuals of the regression sample, respectively. The last model uses a dummy indicating whether a migrant lives in a region further away from its MDF than 4.2km. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.6: Reduced form results, language proficiency

Dependent variable:	Δ Usual language is German				Δ German speaking proficiency				Δ German writing proficiency			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Distance	-0.00				0.01				0.03			
	(0.01)				(0.02)				(0.02)			
Distance>Median		-0.03				0.03				0.10*		
		(0.03)				(0.05)				(0.06)		
Distance>75% percentile			0.00				-0.00				0.11*	
			(0.03)				(0.05)				(0.06)	
Distance>4.2km				0.03				0.10				0.09
				(0.06)				(0.10)				(0.14)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	996	996	996	996	987	987	987	987	986	986	986	986
R-sq.	0.07	0.07	0.07	0.07	0.14	0.14	0.14	0.14	0.11	0.11	0.11	0.10

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.4) using changes between the post and pre-broadband values of German proficiency indicators as dependent variables. For each dependent variable, the first model is estimated using the distance to a municipality's MDF as continuous variable. The second and third models use a dummy indicating whether a migrant lives in a region further away from its MDF than 50% or 75% of individuals of the regression sample, respectively. The last model uses a dummy indicating whether a migrant lives in a region further away from its MDF than 4.2km. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.7: Reduced form results, social integration

Dependent variable:	Δ Frequency of voluntary activity				Δ Received Germans				Δ Visited Germans				Δ Has German friends			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Regressors:																
Distance	(0.01)				(0.02)				(0.02)				(0.02)			
Distance>Median		-0.04				0.01				-0.03				-0.04		
		(0.04)				(0.05)				(0.05)				(0.04)		
Distance>75% percentile			-0.04				0.03				0.04				0.04	
			(0.04)				(0.04)				(0.05)				(0.05)	
Distance>4.2km				-0.00				-0.08				0.02				0.10
				(0.09)				(0.08)				(0.10)				(0.08)
Fixed-effects																
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	1443	1443	1443	1443	796	796	796	796	797	797	797	797	1488	1488	1488	1488
R-sq.	0.09	0.09	0.09	0.09	0.16	0.16	0.16	0.16	0.15	0.15	0.15	0.15	0.10	0.10	0.10	0.10

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.4) using changes between the post and pre-broadband values of social integration indicators as dependent variables. For each dependent variable, the first model is estimated using the distance to a municipality's MDF as continuous variable. The second and third models use a dummy indicating whether a migrant lives in a region further away from its MDF than 50% or 75% of individuals of the regression sample, respectively. The last model uses a dummy indicating whether a migrant lives in a region further away from its MDF than 4.2km. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.8: Reduced form results, other integration indicators

Dependent variable:	Δ Degree of feeling German				Δ Wish to remain in Germany				Δ Life satisfaction			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Distance	0.03				-0.00				0.10***			
	(0.04)				(0.01)				(0.04)			
Distance>Median		0.00				0.01				0.09		
		(0.11)				(0.03)				(0.09)		
Distance>75% percentile			0.08				-0.03				0.07	
			(0.12)				(0.03)				(0.10)	
Distance>4.2km				0.47***				-0.08				0.61***
				(0.18)				(0.06)				(0.23)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	706	706	706	706	1265	1265	1265	1265	1630	1630	1630	1630
R-sq.	0.14	0.14	0.14	0.15	0.12	0.12	0.12	0.12	0.08	0.08	0.07	0.08

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.4) using changes between the post and pre-broadband values of miscellaneous integration indicators as dependent variables. For each dependent variable, the first model is estimated using the distance to a municipality's MDF as continuous variable. The second and third models use a dummy indicating whether a migrant lives in a region further away from its MDF than 50% or 75% of individuals of the regression sample, respectively. The last model uses a dummy indicating whether a migrant lives in a region further away from its MDF than 4.2km. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

2.C.3. Results of placebo regressions using median threshold

Table 2.C.9: Placebo estimation results, economic integration

Distance to MDF:	$\leq 50\%$ percentile						$> 50\%$ percentile					
Dependent variable:	Δ Employment status		Δ Log wage		Change job		Δ Employment status		Δ Log wage		Change job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	0.08**		0.12		0.02		0.02		0.01		-0.01	
	(0.04)		(0.08)		(0.04)		(0.03)		(0.07)		(0.04)	
Δ Broadband use frequency		0.02		0.13		-0.01		0.05		0.02		0.03
		(0.04)		(0.08)		(0.04)		(0.03)		(0.06)		(0.04)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	638	636	329	320	638	636	628	625	317	321	628	625

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of economic integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 50% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.10: Placebo estimation results, language proficiency

Distance to MDF:	$\leq 50\%$ percentile						$> 50\%$ percentile					
Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency		Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	0.07*		0.06		0.09		0.05		-0.09		0.02	
	(0.04)		(0.08)		(0.08)		(0.04)		(0.07)		(0.09)	
Δ Broadband use frequency		0.00		-0.08		0.00		-0.01		-0.05		-0.03
		(0.03)		(0.08)		(0.10)		(0.04)		(0.08)		(0.09)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	430	429	425	424	425	424	421	419	418	416	418	416

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of German proficiency indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 50% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.11: Placebo estimation results, social integration

Distance to MDF:	$\leq 50\%$ percentile								$> 50\%$ percentile							
Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends		Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Regressors:																
Δ Broadband access	0.07		0.05		0.01		-0.06		0.04		0.01		0.01		-0.12**	
	(0.05)		(0.05)		(0.06)		(0.06)		(0.06)		(0.06)		(0.06)		(0.06)	
Δ Broadband use frequency		0.13**		-0.02		-0.03		-0.00		0.02		0.01		0.07		-0.13**
		(0.06)		(0.05)		(0.06)		(0.06)		(0.06)		(0.06)		(0.06)		(0.06)
Fixed-effects																
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	628	626	334	332	334	333	589	587	621	618	333	331	334	331	580	577

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of social integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 50% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.C.12: Placebo estimation results, other integration indicators

Distance to MDF:	$\leq 50\%$ percentile						$> 50\%$ percentile					
Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction		Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Regressors:												
Δ Broadband access	-0.01		0.01		0.28*		0.36***		0.06		0.11	
	(0.18)		(0.04)		(0.14)		(0.13)		(0.05)		(0.15)	
Δ Broadband use frequency		0.02		0.02		0.21		0.41***		-0.02		0.19
		(0.19)		(0.03)		(0.15)		(0.13)		(0.04)		(0.13)
Fixed-effects												
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	341	340	488	486	638	636	340	339	490	488	628	625

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the post and pre-broadband values of miscellaneous integration indicators as dependent variables. Regressions for models (1) to (6) are performed on individuals living in municipalities within the 50% percentile of the distance to their MDFs, models (7) to (12) are based on the remaining 25% of individuals. Controls include changes of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population, female population share, share of working-age population, share of population older than 65 and unemployment rate in the municipality of residence. Standard errors are clustered at the municipality level.

2.D. Appendix to Section 2.8

Table 2.D.1: Placebo estimation results, economic integration

Dependent variable:	Δ Employment status		Δ Log wage		Changed job	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	0.05		-0.05		0.01	
	(0.04)		(0.07)		(0.03)	
Δ Broadband use frequency		0.10**		0.03		0.07
		(0.04)		(0.07)		(0.04)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	673	669	363	360	673	669
R-sq.	0.16	0.17	0.25	0.25	0.18	0.19

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the averages of 1992 and 1993 and pre-broadband values of economic integration indicators as dependent variables. The main regressors still refer to changes between the post and the pre-broadband period. Controls include changes between the averages of 1992 and 1993 and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.D.2: Placebo estimation results results, language proficiency

Dependent variable:	Δ Usual language is German		Δ German speaking proficiency		Δ German writing proficiency	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	-0.08**		0.07		0.07	
	(0.04)		(0.08)		(0.10)	
Δ Broadband use frequency		-0.01		0.21**		0.21*
		(0.04)		(0.11)		(0.12)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	527	525	431	429	429	427
R-sq.	0.15	0.14	0.20	0.20	0.14	0.14

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the averages of 1992 and 1993 and pre-broadband values of German proficiency indicators as dependent variables. The main regressors still refer to changes between the post and the pre-broadband period. Controls include changes between the averages of 1992 and 1993 and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.D.3: Placebo estimation results, social integration

Dependent variable:	Δ Frequency of voluntary activity		Δ Received Germans		Δ Visited Germans		Δ Has German friends	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Regressors:								
Δ Broadband access	0.02		0.03		-0.00		0.12**	
	(0.06)		(0.04)		(0.04)		(0.05)	
Δ Broadband use frequency		0.00		0.03		0.09*		0.08
		(0.07)		(0.04)		(0.05)		(0.06)
Fixed-effects								
Cohort	✓	✓	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓	✓	✓
Obs.	669	665	430	428	431	429	640	636
R-sq.	0.09	0.09	0.15	0.15	0.12	0.13	0.13	0.13

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the averages of 1992 and 1993 and pre-broadband values of social integration indicators as dependent variables. The main regressors still refer to changes between the post and the pre-broadband period. Controls include changes between the averages of 1992 and 1993 and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Table 2.D.4: Placebo estimation results, other integration indicators

Dependent variable:	Δ Degree of feeling German		Δ Wish to remain in Germany		Δ Life satisfaction	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Regressors:						
Δ Broadband access	-0.01		-0.06		0.19	
	(0.11)		(0.05)		(0.15)	
Δ Broadband use frequency		-0.10		-0.03		0.04
		(0.17)		(0.06)		(0.15)
Fixed-effects						
Cohort	✓	✓	✓	✓	✓	✓
Mun. type	✓	✓	✓	✓	✓	✓
Federal state	✓	✓	✓	✓	✓	✓
Origin country	✓	✓	✓	✓	✓	✓
Obs.	426	424	526	523	673	669
R-sq.	0.13	0.13	0.16	0.16	0.11	0.11

* p<0.1, ** p<0.05, *** p<0.01

Notes: This table shows results of regression equation (2.1) using changes between the averages of 1992 and 1993 and pre-broadband values of miscellaneous integration indicators as dependent variables. The main regressors still refer to changes between the post and the pre-broadband period. Controls include changes between the averages of 1992 and 1993 and pre-broadband values of the following variables: Age and years since migration, each as square and cubic, years of education, children under 16 in the household, population and the female population share in the municipality of residence. Standard errors are clustered at the municipality level.

Chapter 3

POLITICAL ESPIONAGE IN THE GERMAN DEMOCRATIC REPUBLIC

3.1. Introduction

In this paper, we provide a comprehensive overview of the political espionage activities of the German Democratic Republic (GDR, more colloquially also referred to as East Germany). We have access to a digital copy of the East German foreign intelligence agency's (*Hauptverwaltung A*, HVA) database, the so-called SIRA data which was spared from the agency's deliberate efforts to destroy any compromising material before the reunification of Germany by a stroke of luck. SIRA contains meta data on a large fraction of all material the East German secret service received from its spies located in the West during the period between 1969 and 1987. It also includes data on all outgoing information that was compiled based on these materials. Although most original incoming and outgoing material was destroyed, the available meta data allow us to infer various aspects of espionage activity.

For each piece of information gathered by a spy between 1969 and 1987, we observe when it was provided to the intelligence agency, some basic description of its form (for example, whether it was a written report or a copy of original documents), and how it was rated in terms of relevance and credibility by the HVA. Importantly, some meta features were explicitly intended to capture the content of the respective information. These include a set of keywords, intuitively similar to subject headers, the name of the institution or organization, and the country to which an information was related. The data also contain some basic features of the spies that were responsible for collecting the respective pieces of information. We observe their unique identifiers, code names

and the agency's assessment of their reliability. For outgoing information, instead, we observe the names of their respective recipients, often high-ranking politicians.

The vast majority of the roughly 160,000 incoming pieces of information were either copies of original materials (40%) or reports written by the spy or her case officer (60%). 25% of the approximately 22,000 outgoing pieces consisted of these reports, the remaining 75% were analyses compiled by dedicated HVA evaluators. Unsurprisingly, most pieces of information were marked as referring to Central Europe or Germany. With respect to institutions, most pieces of information were associated with the NATO, the European Community and West German political parties. The majority of incoming pieces (70%) were rated between *medium value* and *low value* in terms of relevance. For outgoing information, however, most pieces (50%) were rated either as *valuable* or of *medium value*. Only about 1% of incoming and outgoing information was rated as either *very valuable* or *valuable* by the HVA.

The most productive spy, code name JACK, provided 2,177 pieces of information during his 14 years of active service. On average, however, the roughly 6,000 spies we are able to identify in the data, collected significantly less pieces of information, roughly 12, and were actively engaged in espionage for only 2 years.

Most outgoing information was sent to the East German foreign ministry which received approximately 6,000 pieces. On average, one of the 134 distinct registered recipients received 33 pieces of information per year in the course of 4 years.

To document the scope of the data and possibilities for analysis, we implement two case studies in which we filter the meta data for relevant information about two series of negotiations between high-ranking East and West German representatives: the negotiations on the so-called cultural agreement over the years 1973 to 1986, and the negotiations on two loans that were issued to the GDR in 1983. In both cases, we show that the inflow and outflow of related pieces of information increased significantly around the negotiation dates, suggesting that the HVA's espionage activities were explicitly targeted towards high-profile political negotiations.

In a second analysis, we use the keywords provided in SIRA to perform a quantitative analysis on the thematic emphasis of East German foreign intelligence efforts during the period from 1969 to 1987. We fit a Latent Dirichlet Allocation (LDA) model which uncovers 11 key topics of espionage activities. The most important ones, as judged by the number of related pieces of information, we label as *Trade & economic relation with West Germany*, *Political landscape in West Germany* and *Geo & security policy*. Between 1969 and 1987, more than 35,000 incoming and 4,000 outgoing pieces of information were related to these topics. While the yearly share of incoming and outgoing pieces of information related to the topic *Trade & economic relation with*

West Germany was dropping from approximately 27% and 19%, respectively, in 1969 to roughly 12% in 1987, the yearly share for pieces covering the topic *Geo & security policy* increased from 2% and 3%, respectively, to roughly 20% in the same period. In terms of relevance, however, the topic *Military strategy & technology*, scored significantly better: the average relevance score given to pieces related to this topic was 2.3 whereas the other topics were rated with values roughly between 2.6 and 2.7 by the HVA (lower values imply higher relevance). Other topics were related to civil R&D, military technology and activities of enemy individuals, groups or nations.

The existing research on the activities of the East German secret service, the so-called Stasi, and its foreign intelligence department, the HVA, often focuses on specific case studies, individual spies or the inner workings of the agency (Müller-Enbergs, 2001; Macrakis, Friis, and Müller-Enbergs, 2009). The dedicated Stasi Records Archive (*Bundesbeauftragter für die Stasi-Unterlagen*, BStU), which merged with the German Federal Archives in 2021, also provides a comprehensive handbook series (BStU, n.d.) on different aspects of the agency. Using the same data as we do, BStU (2013a) conducted an in-depth analysis of how the Stasi's and the HVA's political espionage activities focused on West German parliaments, parties and MPs.

In economics, Jacob and Tyrell (2010), Friehe, Pannenberg, and Wedow (2015) and Lichter, Löffler, and Siegloch (2021) relate the number of spies which had been located in East German counties to contemporaneous individual personality traits, indicators of social capital and economic outcomes. Glitz and Meyersson (2020) analyze industry-level data on the extent of intelligence provided to the Stasi's spies in West German companies and show that economic espionage of the GDR led to a significantly smaller gap in total factor productivity between both countries.

The remainder of this paper is structured as follows: Section 3.2 describes the historical background into which this research is embedded. It provides an overview of the Stasi and its foreign intelligence agency HVA, its spies and operational activities. Section 3.3 describes the data and provides a general overview of the distribution of information inflows and outflows over time, statistics on the form in which pieces of information were received, the countries and institutions they are referring to and their relevance in the eyes of the Stasi. Section 3.3 also elaborates on how incoming information was used as input for outgoing pieces and gives some details on the spies collecting and the recipients receiving information. Section 3.4 presents the case studies concerned with the inner-German negotiations on a cultural agreement and the loan issued to the GDR and backed by West Germany. Section 3.5 presents the results of the topic model and details on the necessary data pre-processing and model selection exercises. Section 3.6 concludes.

3.2. Historical Background

3.2.1. The East German secret service

In East Germany, espionage activities were primarily carried out by the Ministry for State Security (*Ministerium für Staatssicherheit*, MfS, or, more commonly, just Stasi). However, the activities and responsibilities of the Stasi were not limited to those of a "classical" intelligence agency. The ministry is commonly characterized as a secret police with its own law enforcement personnel, investigators, detention centers and even judges and prosecutors. Its main goal was to ensure the status of the Socialist Unity Party of Germany (*Sozialistische Einheitspartei Deutschlands*, SED) as the only leading political party of the GDR. In achieving this goal, the Stasi applied several methods ranging from surveillance (e.g. phone tapping, wire tapping, video surveillance, controlling correspondence and movement of targets) over imprisonment to *Zersetzung* (probably best translated as degradation) which meant the systematic manipulation and destruction of the target's personality.[1] The Stasi has also been linked to assassinations and terror attacks (Schmole, 2011). To become a target, it was sufficient to be considered a political enemy of the SED.[2]

Foreign intelligence Within the MfS, the *Hauptverwaltung A* (HVA)[3] represented the foreign intelligence unit. While there were other departments concerned with foreign countries, like the Main Departments (*Hauptabteilung*) II (counterintelligence), XX/5 (defense against political-ideological diversion) and XX (monitoring of the economy), it was the HVA which was mandated with aggressive activities on foreign soil. To prepare these activities, the HVA, with the help of its spies, collected vast amounts of potentially relevant information through economic, political and military espionage. Measured by the number of most yielding information sources (BStU, 2013b), economic espionage was the most important of these areas (39%), closely followed by political espionage (38%), our area of focus for this paper. Only 13% of the information sources were tied to military espionage.

[1] Through its own university (Förster, 1996), the MfS provided an almost theoretical-academic framework on various espionage techniques, e.g. on degradation measures which included, amongst others, systematic destruction of a target's public reputation or systematic organisation of occupational and social failures.

[2] For a brief overview of the Stasi, see Dümmel and Piepenschneider (2014). For a comprehensive overview, compare BStU (n.d.). This section is based on information from both sources.

[3] *Hauptverwaltung A* could roughly be translated as Main Administration Department A. Often, but incorrectly (BStU, 2013b), the letter A is taken to be short for *Aufklärung*, the German word for intelligence.

The spies and political espionage In line with the goals of the MfS, political espionage of the HVA was mainly intended to identify and mitigate potential risks for the status quo of the SED as leading party of the GDR. To achieve this goal, the HVA relied on the Stasi's extensive network of spies, the so-called Inofficial Employees (*Inoffizielle Mitarbeiter*, IM). Before the German reunification, roughly 189,000 registered IMs were located in East Germany and about 3,000 in West Germany. As opposed to the official, full-time, employees of the MfS (*Hauptamtliche Mitarbeiter*), the Stasi recruited spies based on their occupational or social positions such that they were able to provide useful information about the field they were working in or the people they had connections to. After recruitment, the IMs continued their civil employment and were often actively supported by the MfS in their career paths. Spies were usually supporters of the SED regime, enforced collaboration was an exception. Still, for many, remunerations for espionage activities was a main reason for collaboration, though payments were often only moderate. Generally, the spies were mandated to collect information about the sentiment and opinions in the East German population, and to detect attempts and trends targeted against the SED. Sometimes, the work of a spy also included active participation in operative measures like investigations or the previously mentioned degradation. Each IM was assigned to an official Stasi employee, her case officer (*Führungsoffizier*), who was responsible for managing the spy, including, for example, receiving her incoming pieces of information. The results of Jacob and Tyrell (2010), Friehe, Pannenberg, and Wedow (2015) and Lichter, Löffler, and Siegloch (2021) show that the impact of the presence of so many undercover spies on the East German society, even today, cannot be understated.

In this paper, however, we focus on the IMs stationed in West, primarily in West Germany, and the information they provided to the MfS.[4] One example of the political espionage activities of these informants was the systematic infiltration and surveillance of the West German federal parliament, its committees and members. Using the same data sources as we do (however, with more access rights), BStU (2013a) show that, amongst others, spies provided information and assessments from the budget, defense and foreign-affairs committees but also about within-party working groups and about single MPs. These spies were mostly not MPs themselves, but simple party members, journalists or secretaries.

Active Measures Although we do not study the operative activities of the HVA (the so-called Active Measures), for illustrative purposes, we provide some examples of how

[4]Sometimes, also official Stasi employees were stationed in the West. We cannot distinguish whether information were provided by these so-called Special Operations Officers (*Offiziere im besonderen Einsatz*) or "normal" IMs.

the HVA, potentially based on previously received information, actively interfered in West German politics. To support the SED's foreign policy, the MfS was trying to strengthen those political forces in West Germany with, in their eyes, favourable views and weaken those with less favourable attitudes. It is likely that information provided by IMs in West Germany helped the Stasi and SED in identifying West German politicians as members of either group.

One of the most famous cases of operative measures was revealed when the former head of the HVA, Markus Wolf, after the German reunification, claimed that the HVA had paid 50,000 € to Julius Steiner, a conservative member of the West German federal parliament. In return, Steiner voted against his party's candidate Rainer Barzel and, instead, for the current Social Democratic chancellor Willy Brandt during a constructive vote of no confidence in 1972. This vote of no confidence was very important to the SED because it favoured Brandt's stance towards the relationship between East and West Germany and therefore wanted to keep him in power. Buying a vote of even another conservative MP, the MfS had a decisive impact in the failure of the vote of no confidence against Brandt. Ironically, however, in 1974, Brandt resigned because his personal referee, Günter Guillaume, was exposed as an East German spy.

Other examples of so-called active measures included covert operations against West German politicians which the SED deemed hostile. For example, the HVA tried to prevent the election of Konrad Adenauer as West German chancellor in 1957. On top of collecting compromising materials, the HVA also engaged in providing these to the press or even faking documents to fuel tensions among the conservative parties in West Germany (BStU, 2013a).

3.2.2. Historical sources

Next to interviews and biographies of contemporary witnesses, studies on the HVA are based on three different sources, all of which are maintained by the German Federal Archives: first, some files compiled by the Stasi/HVA were not destroyed during its resolution and thus are still accessible.[5] Second, the German government was able to retrieve a copy (saved on microfilm) of personal records of the HVA from the CIA which, in turn, likely received it from a KGB agent in 1988 (Müller-Enbergs, 2007). These so-called *Rosenholz* records contain the clear names of some IMs, official employees and, mostly, other individuals related to the HVA, often sources which were exploited by spies without their knowledge. Finally, our data on political espionage activities of the Stasi stems from the HVA's main electronic database SIRA (*System der Information-*

[5] A lot of original outgoing pieces of information, which mostly took the form of written reports, can still be accessed in the German Federal Archives (BStU, 2013b).

srecherche der Hauptverwaltung Aufklärung). SIRA's subdatabase 12 (*Teildatenbank 12*) contains the records on politically relevant information that the HVA received from its IMs in the West or sent out between the years 1951 and 1989.[6]

Dedicated HVA units were tasked with the process of archiving, evaluating and forwarding all pieces of incoming information relevant for political espionage. An integral part of this process was the creation of an electronic entry in the SIRA database containing, amongst others, the date of arrival of the information, its source, form (written, digital etc.), language, relevance and credibility. To describe the content, the MfS employees used a large set of keywords. Once the information was archived and evaluated, it was passed on to its receivers, mostly entities not directly related to the Stasi like high-renking SED members (Selvage and Süß, 2019) or friendly intelligence agencies such as the KGB (Konopatzky, 2019).

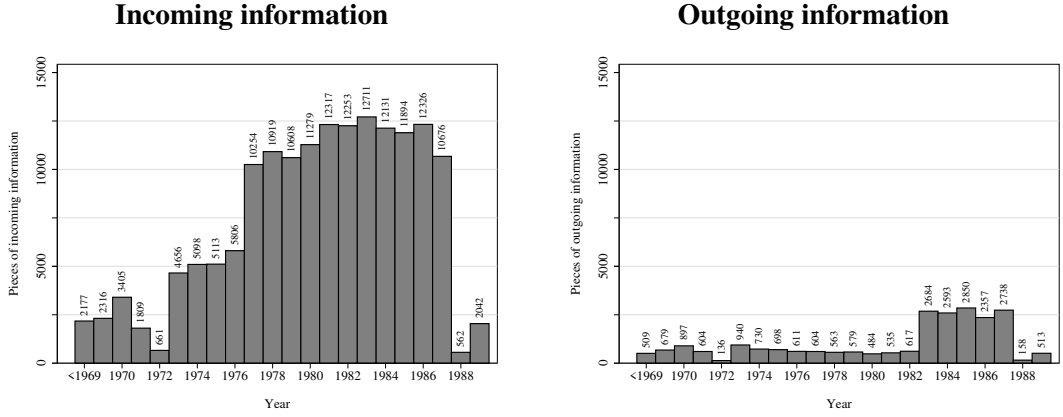
The SIRA data are available today only by a stroke of luck. In the course of German reunification in 1990, the Stasi was disbanded and it was decided to physically destroy all sensitive information, including all electronic data carriers. By 19 March 1990, "10,611 magnetic tapes, 5,267 disks, 544 removable hard disks, and 80 sacks of loose magnet tape material" had been destroyed, including all data stored in the original SIRA system (Engelmann, Halbrock, and Joestel, 2020). However, the HVA had created security copies of the SIRA records because of a plan to conduct a data conversion of the entire system in 1988/1989. These copies were overlooked upon the liquidation of the Stasi. During the 1990s, the BStU reconstructed the data from these copies (Konopatzky, 2019). Although only meta features, the SIRA records constitute an invaluable opportunity for quantitative analyses of East Germany's espionage activities.

3.3. Data

Our excerpt of SIRA's subdatabase 12 contains records on 184,092 distinct pieces of information for each of which we observe whether it was received by the HVA from a spy (incoming information) or sent out by the HVA to a third party (outgoing information). We also observe the respective dates for all pieces of information. Since SIRA was set up by the Stasi in 1969, any record on older pieces of information constitutes a copy of already existing entries in the precursor database. Likely, even the original records before 1969 were incomplete. Furthermore, there is definitely only a fraction of the original records available for the years 1988 and 1989. Our analysis therefore

[6]SIRA also contains other subdatabases, for example subdatabase 11 which is related to economic espionage and analyzed by Glitz and Meyersson (2020).

Figure 3.3.1: Information inflows and outflows between 1969 and 1989



Notes: This figure shows the distribution of the number of incoming and outgoing pieces of information to and from the HVA over years.

focuses on the years between 1969 and the end of 1987.[7]

Within this period, records on 178,131 distinct pieces of information are available, implying an average flow of 9,375 pieces per year. Most of of these, 156,232 (88%), are records on information that was provided by informants to the HVA. The remaining 21,899 records refer to information that was passed on to third parties by the HVA. Figure 3.3.1 displays the inflows and outflows of information from before 1969 to 1989. Besides the very low values in the years before 1969, in 1972, 1988 and 1989, the figure shows a steady increase in the flow of incoming information with a significant jump in 1977, peaking at a value of 12,711 in 1983 and moderately declining thereafter. The number of outgoing pieces of information is much smaller and fluctuates roughly between 500 to 900 pieces per year until 1982. In 1983, the number then increases significantly to 2,684 and changes only very moderately after.

From the meta features we observe (compare Table 3.A.1 in the appendix), we use the following for our analyses: first, descriptions of the form of a piece of information, i.e. whether it constitutes an original or a copy and the data medium on which it was provided to the HVA. Second, we look at the countries and institutions or organizations to which an information can be linked. Third, we study measures of a piece's relevance and level of confidentiality. We then provide an overview how incoming and outgoing pieces are related, for example by showing how many incoming pieces were used to create an outgoing piece of information. We proceed to give more details on the spies who collected information in West Germany and on the receivers of outgoing information. Finally, we analyze the keywords used to summarize the content of pieces of informa-

[7]Given the relatively low number on information flows in 1972, it is likely that for this year the SIRA records are also incomplete (Konopatzky, 2019).

tion. Since original data on incoming information is not available, the keywords are our only source for investigating their content.[8]

From now on, any reference to the total number of incoming or outgoing pieces will refer only to information dated between 1969 and 1987. When explaining the different values of meta features, all percentages and shares mentioned are based on only those pieces of information for which values are non-missing unless explicitly stated otherwise.

3.3.1. Overview of pieces of information

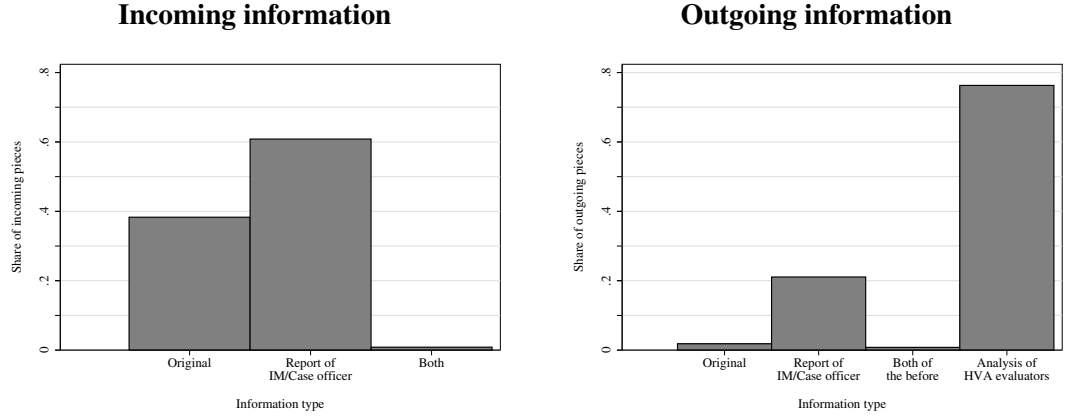
Form For 98% of incoming and 96% of outgoing pieces of information, respectively, SIRA contains a meta feature categorizing each piece according to its form. The vast majority of these pieces, 99%, exhibit exactly one value for this feature. Some pieces, however, fall into multiple, as many as 4, form categories. Still, there are three basic form categories which are almost mutually exclusive and collectively exhaustive. Two of these were written formats: either *Reports*, drafted by the IM or the responsible case officer based on the IM's activities, or *Analyses*, compiled by dedicated HVA evaluators. The latter form of information is only featured among outgoing pieces. The HVA's intern evaluation unit would receive incoming pieces of information and, based on those, write the respective analyses. The last basic information form were *Original Materials* or copies thereof. Figure 3.3.2 plots the respective distributions for incoming and outgoing pieces, respectively. Roughly 40% of incoming information are marked as original materials or copies thereof and almost all remaining pieces were reports. Only a very small fraction consisted of both originals and a report. Among the outgoing information, 75% were analyses, roughly 20% were simply the incoming reports of IMs or their case officers, and the rest consisted of the two remaining categories.[9]

Unfortunately, out of all information marked as originals or copies, only 178 incoming pieces exhibit at least one additional form value providing further insight into the nature of the respective piece. Most of these information are either *Studies* (72), *Reports* (36, i.e. the original information was a report), *Conference Materials* (14), *Working Instructions* (10) or *Protocols* (9). Some other descriptive labels are *Annual Report*, *Technical Documentation*, *Map*, *Phone Register* and *Name Register*.

[8]Due to data protection, we do not have access to 3 meta features: first, the degree of confidentiality of an incoming piece of information at its origin entity. Second, the association of pieces of information with individuals they were referring to. Third, a meta feature which was comprised of non-standardized text entries in which HVA employees further elaborated on the content of an information, often including the exact title of the original piece of information. Especially the last meta feature would have proven very helpful for content analysis.

[9]Note that we are not sure whether multiple form values represent a finer description of the piece of information or indicate multiple source materials.

Figure 3.3.2: Empirical distribution of information form



Notes: This figure shows the distribution of information forms over pieces of information.

To provide an alternative overview of original pieces, we also investigate the meta feature describing the data media as which the information arrived or left the HVA. All but 108 of the 60,084 incoming and one of the 551 outgoing pieces which are marked as originals or copies, have non-missing values for this meta feature. 99.8% and 100% of these pieces, respectively, were marked as *Paper*. Of these, 3,764 incoming and 383 outgoing pieces exhibit one additional data medium value[10] which is *Attachments* for 3,722 incoming and all outgoing pieces. Other values occurring as data media are *Brochures*, *Folders*, *Books* or *Films*, i.e. mostly still written materials.[11] The 0.2% original incoming pieces which are not marked as papers were labelled *Attachments*, *Brochures*, *Books*, *Folders*, *Direction Materials*, *Fotos*, *Film*, *Magnetic Disk* or *Micro Fiche* instead.

On average, incoming pieces of information consisting of original materials and further marked as paper are 32 pages long, with a median value of 14 pages. The respective distribution is very spread out, with a 99% percentile of 294 pages and a maximum value of 6,000 pages. All but 48 of the 93,330 incoming information consisting only of a report also came in paper form. The according page number distribution is considerably more narrow than the previous one: the median piece was 4 pages long, the average 5, the 99% percentile is 38 pages and the longest report 2,265 pages long. From the

[10] With the exception of one incoming piece which exhibits two further values.

[11] Unlike for the information form meta feature, we believe that multiple data medium values per piece of information might indeed indicate distinct source materials and do not represent a finer description of the respective piece. For example, a piece marked as *Paper* and *Book* likely consisted of some papers and, additionally a book. Brochures, folders and books, albeit being made out of paper, in the HVA logic, did not fall into the data medium category *Paper*. For each distinct data medium value per piece of information, SIRA also contains the quantity of the respective medium. For most pieces of information marked with multiple data medium values, one of which was *Paper*, the number of data media varies across data media values indicating different source materials.

15,999 outgoing pieces which are analyses, we observe the data media on which the information was delivered in 74% of cases, all of which are marked as paper. The page number distribution for analyses is the most narrow, most pieces were only one page long which is also the median value. On average, an analysis was 3 pages long, the 99% percentile is at 19 pages, the longest document had 105 pages. Figure 3.A.1 in the appendix plots all three distributions. Obviously, original material in paper form was considerably more extensive than incoming reports which, in turn, were still longer than the analyses of the HVA evaluators. It seems reasonable to believe that Stasi employees (case officers and evaluators) were mandated to condense the content of incoming information before using them as input for outgoing pieces.

Relevant countries and institutions The following paragraph sheds some light on the countries and institutions to which pieces of information can be linked using the values of the respective meta features. For 84% and 77% of incoming and outgoing pieces of information, respectively, SIRA contains values on the countries to which an information is related. Most pieces exhibit multiple country values, the mode of the respective distribution is 4 countries per piece of information, one piece is even labelled by 88 different country descriptors. Table 3.A.3 in the Appendix lists the 20 most frequent country descriptors used for incoming and outgoing pieces, respectively. Among incoming pieces for which a country reference is available, most referred to *Central Europe* (56.5%) and *West Germany* (46.6%) in particular. Following, albeit with roughly a 15 percentage point difference, were information related to *America* and the *USA*. This ranking is similar for outgoing pieces, however the relative importance of information on *Central Europe* and *Germany* seems to have been higher as indicated by the fact that 66.5% and 57.3% of outgoing pieces were labelled with those country references, respectively. 15.7% and 19.8% of incoming and outgoing pieces were tied to the *USSR*, respectively, ranking 6th for both types of information. The 5th place, however, was tied to different countries for incoming and outgoing pieces: while, with 19.4% of pieces labelled, the *Far East* seemed to be a relatively important area for incoming information, it only ranked 8th (with 17.1% of pieces labelled) among outgoing pieces. Instead, the 5th most frequent country label for outgoing information was *GDR*. Next to labels *Europe*, *Japan*, *Iran* and *Northern Africa*, this constituted the highest difference in the ranking of country references between incoming and outgoing pieces.

For roughly 56% and 49% of incoming and outgoing information, respectively, SIRA also contained data on the organization or institution to which a piece was linked. More than half of pieces of information exhibit multiple institutional descriptors, the mode, however, is one organization value per piece. Two pieces of information are tied

Table 3.3.1: Most frequent institutional references

Institution/Organization		Translation/Explanation	% of pieces labelled		
			Incoming	Outgoing	(rank)
1	NATO		15.9	22.0	(1)
2	SPD	Social Democratic Party of Germany	14.0	15.4	(2)
3	EG	*Europäische Gemeinschaft* (European Community)	11.5	11.8	(3)
4	CDU	Christian Democratic Union of Germany (Christian conservative party)	10.2	10.3	(4)
5	FDP	Free Democratic Party (liberal party)	7.3	7.0	(5)
6	CSU	Christian Social Union in Bavaria (Christian conservative party)	5.7	6.9	(6)
7	PLO	Palestine Liberation Organization	5.6	4.3	(10)
8	BUNDESTAG	West German federal parliament	4.2	2.8	(15)
9	UNO		3.7	4.6	(8)
10	BUNDESWEHR	West German armed forces	3.2	6.3	(7)
11	BDI	*Bundesverband der Deutschen Industrie* (Federation of German Industries)	3.1	3.5	(11)
12	GRUENE	The Greens (West German green party)	2.9	2.5	(17)
13	SENAT	State Cabinet of West Berlin	2.6	4.5	(9)
14	WV	N/A	2.3	2.9	(14)
15	DGB	*Deutscher Gewerkschaftsbund* (German Trade Union Confederation)	2.0	3.0	(13)
16	RGW	*Rat für gegenseitige Wirtschaftshilfe* (Socialist pendant of OECD)	1.6	2.7	(16)
17	AA	N/A	1.5	3.2	(12)
18	ABGEORDNETEN-HAUS	State parliament of West Berlin	1.4	0.8	(42)
19	FES	*Friedrich-Ebert-Stiftung* (Friedrich-Ebert-Foundation)[1]	1.2	1.5	(22)
20	SI	N/A	1.1	0.9	(34)
21	SED		1.1	1.6	(20)
22	BMVG	*Bundesministerium der Verteidigung* (Federal Ministry of Defence)	1.1	1.9	(18)
42	WARSCHAUER VERTRAG	Either: Treaty of Warsaw[2] or: Warsaw Pact	0.6	1.6	(19)

Notes: This table shows the percentage of pieces of information labelled by the 20 most frequently used institution descriptors for incoming pieces. The percentage and rank (in parentheses) of the same institution descriptors among outgoing pieces of information is also given. The last 3 institution descriptors are shown because they are among the 20 most frequently used descriptors for outgoing pieces of information. "N/A" in column 2 indicates that we do not know to which organization or institution the descriptor in column 1 refers. **1:** Traditionally, each West German party maintains a foundation which, amongst other purposes, engages in political education and maintaining its party's archives. For legal reasons these foundations are separate legal entities. The Friedrich-Ebert-Foundation is the foundation of the SPD. **2:** The Treaty of Warsaw (signed 1970 and ratified 1972) was intended to normalize the diplomatic relations between West Germany and Poland.

to 25 organizations each. Table 3.3.1 shows the 20 most frequent institutional references for both types of information.

Among both, the ranking for the first 6 organizations is equal: the label *Nato* was used for 15.9% of incoming and 22% of outgoing pieces. Following, were the different

West German parties *SPD*, *CDU*, *FDP* and *CSU* as well as *EG*, likely the abbreviation for the German translation of European Community. The biggest ranking differences occur for institutional labels *Abgeordnetenhaus* (State parliament of West Berlin), *SI* (we do not know the meaning of this abbreviation) and *Warschauer Vertrag* (the Treaty of Warsaw or the Warsaw Pact). The first two of these labels occur more frequently in incoming than in outgoing pieces whereas the last one is used relatively more often for outgoing information.

Relevance Not all the information provided by spies were equally relevant. For each piece of information, the HVA evaluators assigned a relevance score with values 1 (*very valuable*), 2 (*valuable*), 3 (*medium value*), 4 (*low value*) and 5 (*without value*) and stored them in SIRA. This meta feature is available for roughly 92% of incoming pieces of information and almost none, only 89, outgoing pieces. Still, we can get an overview of the relevance of outgoing information by observing the relevance of the incoming pieces serving as their input.[12] Figure 3.3.3 plots the respective distributions. Most incoming pieces, roughly 70%, received a valuation between the median value of 3 and 4.[13] A considerable fraction, 20%, were rated between values 2 and 3, and only very few pieces were rated very valuable or considered as without value.

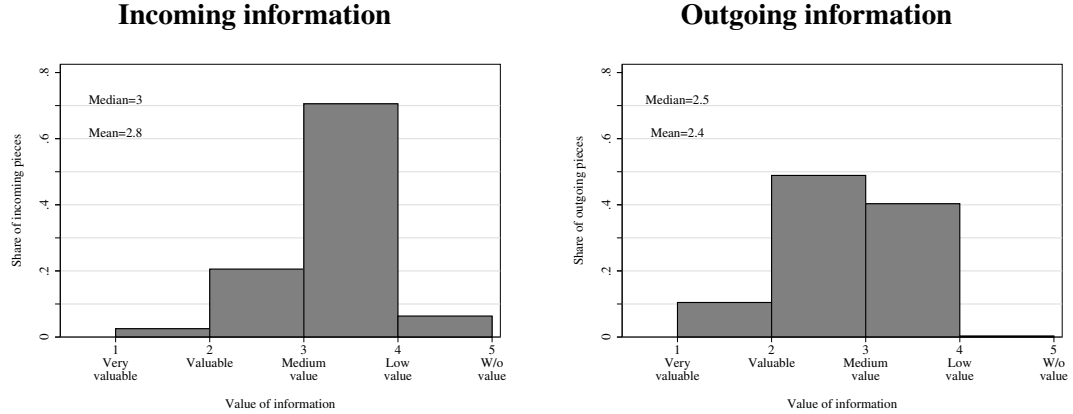
Clearly, the distribution of outgoing pieces of information over value scores exhibits more mass on smaller values, indicating that the HVA used only the most relevant incoming pieces as inputs for creating outgoing information.

Confidentiality The confidentiality of incoming pieces is correlated with their relevance score. The classification of information with respect to confidentiality is available for 76% of incoming but only 1.6% of outgoing pieces. The latter ones were all labelled with the highest degree of confidentiality, that is a value of 1, meaning *Secret Undisclosed Material*. For incoming pieces, three more values were available: 2 (*Confidential Undisclosed Material*), 3 (*Internal Material*) and 4 (*Disclosed Material*). Overall, roughly 17% of incoming pieces were rated as *Secret Undisclosed Material* and almost all remaining pieces as *Confidential Undisclosed Material*. However, among those pieces which contained very valuable and valuable information, more than 30% were rated as *Secret Undisclosed Material*. Across all valuation scores, almost none of

[12] From late 1982 onward, for outgoing pieces of information, HVA employees recorded the unique identifier of each piece of information that served as input for an outgoing piece of information. For 75% of the 21,899 outgoing pieces of information in the relevant period, we observe identifiers of input information and are able to merge 11,093 of these with relevance scores. We elaborate on the link between outgoing and incoming pieces of information later in this subsection.

[13] Since the values sometimes vary per piece of information, the figure shows average values per piece of information.

Figure 3.3.3: Empirical distribution of relevance of pieces of information



Notes: This figure shows the distribution of relevance of pieces of information. Since the values vary per piece of information, the figure shows average values per piece of information. The value of outgoing pieces of information was mostly not assessed by the HVA, instead the right panel is based on the average value of incoming pieces of information used as input for outgoing pieces.

the pieces were rated as *Internal* or *Disclosed Material*.

3.3.2. Linking pieces of information

From late 1982 onward, for each outgoing information, SIRA lists the unique identifiers of those pieces of information which served as input for the respective outgoing piece.[14] Out of the 13,222 outgoing pieces compiled in this period, 86% (11,332), have non-missing values in the meta feature containing their input information. The remaining 14% of pieces of information are either not based on any input or the respective value is indeed missing. In total, 19,622 distinct pieces are listed as inputs for outgoing information, only 9 of them being outgoing pieces of information themselves.[15] This means that out of the 59,738 incoming pieces registered in SIRA between 1983 and 1987, roughly 38% were used to create outgoing information.[16]

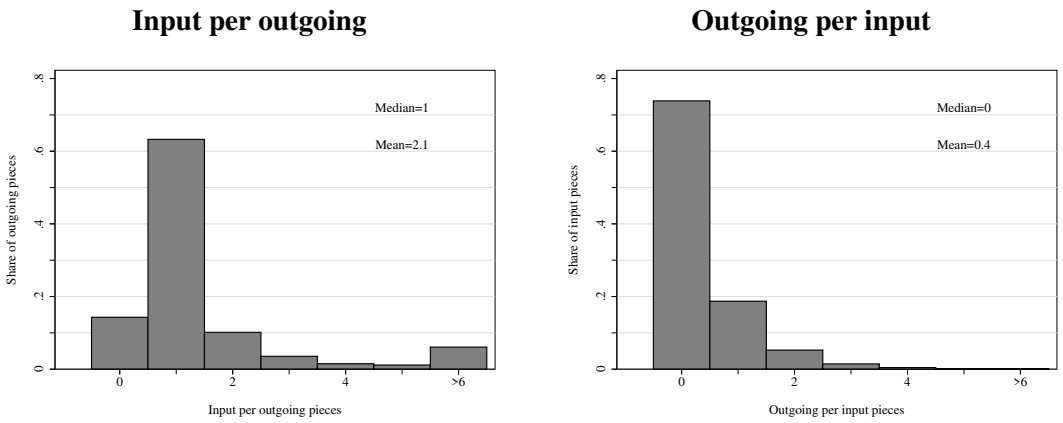
Assuming that missing values imply an outgoing piece was created without an explicit input information, each outgoing information, on average, is based on 2.1 input

[14]Before November 1982, the respective meta feature only contains a running number allocated to each piece of information on top of its unique identifier. Since this number is reset yearly it does not uniquely identify pieces of information and prevents us from relating incoming and outgoing pieces before late 1982. Further, for simplicity, we drop the last months of 1982 for which we could link incoming and outgoing pieces. These are only 49 pieces.

[15]We use the term input pieces instead of incoming pieces to make explicit that, albeit in very few cases, outgoing information can also be used as an input for other outgoing pieces.

[16]Note, however, that 453 incoming pieces used to create outgoing information between 1983 and 1987 were received by the HVA before 1983. Also, 96 of these input pieces are, for unknown reasons, not included in our data. 94 of those have an identifier that matches the form of all other identifiers, 2 are listed with identifiers of a different form.

Figure 3.3.4: Empirical distribution of input per outgoing and outgoing per input pieces

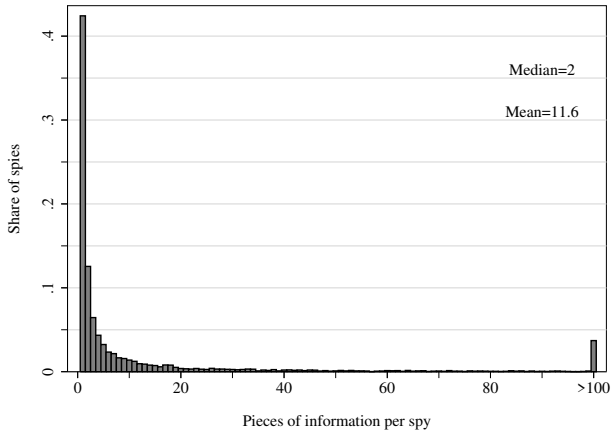


Notes: The left panel of this figure shows the distribution of the number of pieces of information used as input for outgoing pieces of information. Zero input pieces per outgoing piece could mean that the respective outgoing information are not based on any particular input piece or that observations are missing. The right panel of this figure shows the distribution of the number of outgoing pieces of information created by using one input piece.

pieces whereas the median outgoing piece was based on only one input piece. Excluding the missing values, the mean is 2.4 and the median value remains unchanged. The maximum number of input pieces used for a single outgoing information was 146. The left panel of Figure 3.3.4 plots the respective distribution. Note that the means are higher than obtained by dividing the 19,622 distinct input pieces by the 13,222 (or 11,332) outgoing pieces because a piece of information can be used as input for multiple outgoing pieces. As already mentioned, only a minority of incoming pieces of information registered between 1983 and 1987 (and almost none of the outgoing ones) were used as inputs for outgoing information. On average, each incoming information was only used 0.4 times and at most 29 times. The right panel of Figure 3.3.4 plots the respective distribution.

Incoming pieces used as inputs for outgoing pieces are somewhat different than other incoming pieces. Figures 3.A.2 and 3.A.3 in the appendix compare the distribution of information form and relevance between those incoming pieces used and not used as inputs. To make the plots comparable, they are all based on information dated between 1983 and 1987. The share of original materials among incoming pieces used as inputs, 40%, is significantly higher than for information not used as input, 32%. The difference in relevance scores is also striking: although the median piece of information was rated with a medium value for input and non-input pieces alike, the respective distribution for input pieces places more mass on better valuations. The average value of incoming information used as input is 2.5 whereas it is 3 for the other pieces.

Figure 3.3.5: Empirical distribution of the number of incoming pieces of information over spies



Notes: This figure shows the distribution of the number of incoming pieces of information over spies. For better readability, all instances of spies providing more than 100 pieces of information are shown in the rightmost bin. Based on incoming information only.

3.3.3. Informants

For 76% of incoming pieces, the SIRA records contain data on the spy who collected the respective information.[17] Based on their registration numbers and code names, we can identify 5,983 distinct IMs providing information to the HVA from 1969 to 1987. On average, these spies were active for two years and sent roughly 12 pieces of information to the Stasi. The median informant, however, produced only two pieces during her active period which was less than a year. Consequently, the distributions of collected pieces of information and active periods over spies are both right-skewed with most informants providing only few pieces during a short period and a handful of long-term spies sending many pieces.

Figure 3.3.5 plots the empirical distribution of the number of pieces of information per spy. Table 3.3.2 gives an overview of the right end of that distribution and lists the 20 most productive informants within the relevant period along with other relevant meta features. Regarding the number of pieces of information collected, these spies were exceptional. In terms of reliability, however, they did not outperform the majority of the other registered IMs. The left panel of Figure 3.3.6 plots the empirical distribution of reliability scores assigned to spies by the HVA. Most spies (and the median spy) were graded with an A, which, according to the HVA definition, indicated a *reliable* IM. Further reliability scores include values B (*trustworthy*), C (*not checked*), D (*questionable*)

[17] For 24% of outgoing pieces these data are also available, however they refer to the spies collecting the input information on which the respective outgoing piece is based on. Therefore, we only focus on incoming pieces of information in this section.

Table 3.3.2: Top 20 Informants, 1969 - 1987

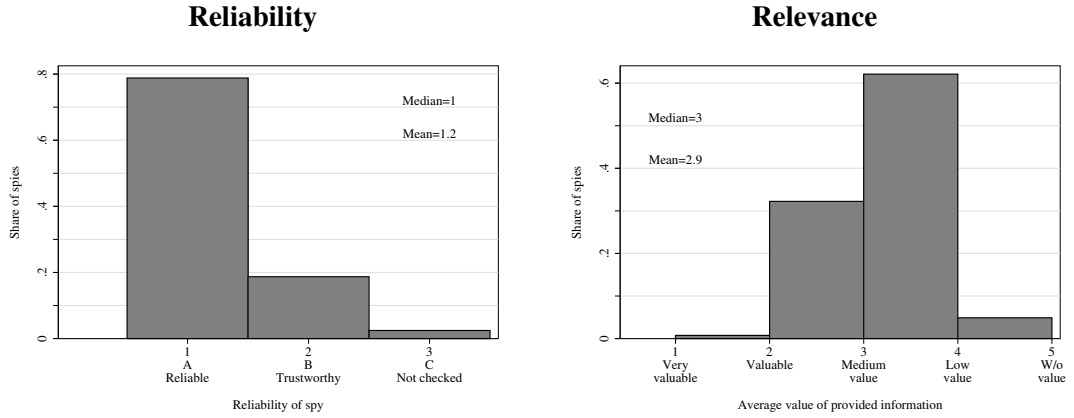
Code name	# Pieces	Reliability	Relevance	Most frequent institutional reference[1]	Active [Years]		
					First	Last	Period
JACK	2177	A	2.7	BDI, EG, U-VERBAND	’73	’87	14
FRIEDRICH	1930	A	2.8	CDU, SPD, FDP	’78	’87	9
AHMED	1755	A	2.2	PLO, EG, UNO	’78	’87	9
FICHTEL	1738	A	2.9	CDU, SPD, EG	’69	’86	17
GERALD	1589	A	2.2	NATO, EG, KSZE	’74	’87	13
GERHARD	1465	A	2.7	NATO, CNAD, BUNDESWEHR	’76	’87	11
PETER	1450	A	1.7	BUNDESWEHR, WV, NATO	’72	’87	15
MAX	1435	A	2.4	SPD, CDU, FDP	’69	’87	18
CLAUS	1370	A	2.8	BDI, EG, U-VERBAND	’69	’85	16
ERICH	1241	A	2.5	NATO, CNAD, DRG	’69	’87	18
ROEDEL	1158	A	2.0	NATO, CNAD, NAFAG	’70	’87	17
MERTEN	1043	A	2.2	EG, NATO, AA	’69	’87	18
JUTTA	1026	A	2.8	EG, SENAT, BUNDESTAG	’69	’87	18
TOPAS	1009	A	2.1	NATO, WV, DPC	’75	’87	12
NORBERT	998	A	2.4	EG, UNO, WELTBANK	’71	’85	14
ADLER	962	A	2.2	EG, AA, UNO	’69	’87	18
REINHARD	922	A	2.1	PLO, NATO, WV	’80	’87	7
BOB	921	A	2.7	SPD, FES, CDU	’71	’87	16
HANS	910	A	2.7	SPD, SENAT, ABGEORDNETENHAUS	’69	’87	18
BERGER	843	A	2.9	EG, BMFT, AA	’69	’87	18

Notes: This table shows the 20 informants who gathered most information between 1969 and 1987. Since reliability scores vary at the piece of information level, the respective column is based on the mode (across all pieces of information gathered by a spy) of the recorded assessments. The HVA rated reliability on an ordinal scale with values A (*Reliable*), B (*Trustworthy*), C (*Not checked*), D (*Questionable*) and E (*Double agent*). Relevance is measured by the average (across all pieces of information gathered by a spy) of the recorded assessments. The HVA rated relevance on an ordinal scale with values 1 (*Very valuable*), 2 (*Valuable*), 3 (*Medium value*), 4 (*Low value*) and 5 (*Without value*). Spies are identified by registration numbers which we do not show here. Instead, we show code names. One spy can have different code names, therefore the code name shown is the mode of the recorded names. Additionally, the three most occurring institutional references among all pieces collected by a spy are given. Based on incoming information only. **1:** See Table 3.A.4 in the appendix for further details on the listed institutions.

and E (*double agent*). Only values A, B and C appear in the data. Encoding this scale with values 1 (A) to 3 (C) results in an average reliability score of 1.2 and a median of 1.[18]

[18] Although reliability, intuitively, should vary at the spy level, it actually varies at the piece of information level (most likely because the reliability of spies was re-evaluated sometimes). Therefore, the histogram in the left panel of Figure 3.3.6 is based on the mode (across all pieces of information gathered by a spy) of the recorded reliability assessments. Encoding the ordinal reliability scale provided in SIRA and taking averages across all pieces of information gathered by a spy barely changes results (compare Figure 3.A.4 in the appendix).

Figure 3.3.6: Empirical distribution of spies' reliability and relevance of provided information



Notes: The left panel of this figure shows the distribution of reliability scores over spies. The HVA rated reliability on an ordinal scale with values A (*Reliable*), B (*Trustworthy*), C (*Not checked*), D (*Questionable*) and E (*Double agent*). Since reliability scores vary at the piece of information level, the plot is based on the mode (across all pieces of information gathered by a spy) of the recorded assessments. The right panel of this figure shows the distribution of the average (across all pieces of information gathered by a spy) relevance scores over spies. The HVA rated relevance on an ordinal scale with values 1 (*Very valuable*), 2 (*Valuable*), 3 (*Medium value*), 4 (*Low value*) and 5 (*Without value*). Based on incoming information only.

In terms of the relevance of the provided information, the most productive spies scored better than other IMs. The right panel of Figure 3.3.6 plots the distribution of the spies' average relevance score (across all pieces of gathered information). The average, and median, IM provided pieces of information of medium value (3) whereas the most productive spies' average relevance scores are often below these values. Especially IM PETER, with an average score of 1.7, did not only provide plenty but also very useful information. This is truly exceptional because among those few spies who had an average relevance score below 2 (43 individuals), most provided considerably less than 100 pieces of information in their active period (often between only 1 and 3 pieces). According to Müller-Enbergs (2007), IM PETER was stationed at the *Bundesnachrichtendienst* (BND), the West German foreign intelligence agency. Using the meta feature on institutions and organizations, we find that most pieces of information collected by IM PETER were associated with the West German military, the Warsaw Pact or Treaty of Warsaw, the NATO (all three shown in column five of Table 3.3.2), the BND and the PLO. These considerations give a first glimpse of what was considered valuable information by the HVA.

The exceptional amount of pieces of information gathered by the spies listed in Table **??** also reflects the fact that they were active for a significantly longer period than the average informant. The top panel of Figure 3.A.5 in the appendix shows that, for the

median spy, the time span between her first and last information provided within the relevant period was less than a year. On average, this period lasted 2.3 years. These numbers are well below those reported for the informants listed in Table 3.3.2. The spikes in the lower panels of Figure 3.A.5 around 1970 for first active years and in 1987 for the last active year are likely caused by the fact that data before and after those dates is missing to a large extent.

3.3.4. Recipients

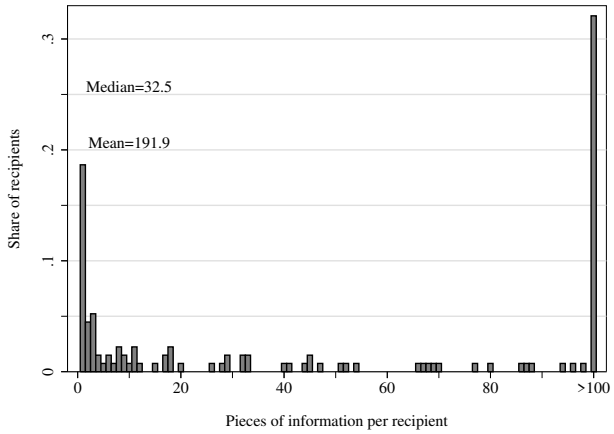
For outgoing pieces created after 1981, HVA employees saved their respective recipients in the SIRA database. This meta feature is non-missing for 48% of all outgoing pieces.[19] Based on their names, we can identify 134 distinct recipients receiving information from the HVA from 1981 to 1987. On average, these recipients were provided with information for three years and received roughly 192 pieces from the Stasi. The median recipient, however, received only 33 pieces over the course of 4 years. As was the case for informants, the distributions of received pieces of information over recipients is right-skewed with most recipients receiving only few pieces and a handful being provided with many pieces. Figure 3.3.7 plots the empirical distribution of the number of pieces of information per recipient and Table 3.3.3 gives an overview of the 20 recipients which received most pieces of information during the relevant period along with other relevant meta features.

In terms of the relevance of the received information, the top 20 receivers scored marginally better than their peers. Figure 3.A.6 in the appendix plots the distribution of a recipient's average (across all pieces of information received) relevance score. Since we are only considering outgoing pieces, these relevance scores are based on the input pieces used to create the respective outgoing pieces. The average, and median, recipient received pieces of information which were both rated with a score of 2.4. The top 20 recipients' average relevance score is 2.2.

Table 3.3.3 shows that recipient MFAA, the East German ministry for foreign affairs, with roughly 6,000 pieces, received considerably more pieces of information than the other recipients. Following, with 1,600 pieces is the MAH, the *Ministerium für Innerdeutschen Handel, Außenhandel und Materialversorgung*, probably best translated as Ministry for Trade. Ranks 3, 4, 5 and 6 belong to high-ranking SED party members, Hermann Axen, Oskar Fischer, Günter Sieber and Erich Honecker, the latter being the de-facto leader of the GDR from 1971 onward. Descriptors IPW, ZK and NVA stand for

[19] 99% of incoming pieces also have non-missing values for this meta feature. For these pieces, however, the values record the HVA department which was responsible for evaluating the respective incoming piece. Therefore, we only focus on outgoing pieces of information in this section.

Figure 3.3.7: Empirical distribution of the number of outgoing pieces of information over recipients



Notes: This figure shows the distribution of the number of outgoing pieces of information over recipients. For better readability, all instances of recipients receiving more than 100 pieces of information are shown in the rightmost bin. Based on outgoing information only.

Institut für Internationale Politik und Wirtschaft, a type of East German think-tank, *Zentralkomitee*, the most important SED executive committee, and *Nationale Volksarmee*, the East German armed forces. All other names belong to high-ranking SED party members.

3.3.5. Keywords

In this section, we provide an overview of the most important meta feature regarding the content of a piece of information. For virtually all incoming and outgoing pieces, we observe one or more keywords describing the content of the respective information. From 1969 to 1987, pieces of information were labelled with 1,095 distinct and a total of 1,305,323 keywords. For incoming information 1,073 distinct and a total of 1,139,800 descriptors were used, outgoing pieces are associated with 1,000 distinct and 165,523 total words. The distribution of the number of keywords over pieces of information is skewed to the right for both incoming and outgoing information. On average, each piece of incoming information is described by 7.3 and each piece of outgoing information by 7.6 distinct keywords while the respective median values are 7 and 6, respectively. 1,374 incoming but only 2 outgoing pieces are not associated with any keyword.[20] The maximum number of keywords per information is 70 among incoming and 66 among outgoing pieces, both values occurring only for one piece of informa-

[20]We do not know whether these are missing values or, indeed, pieces not associated with any keyword. In this section and all analyses based on keywords, we are assuming the latter.

Table 3.3.3: Top 20 Recipients, 1969 - 1987

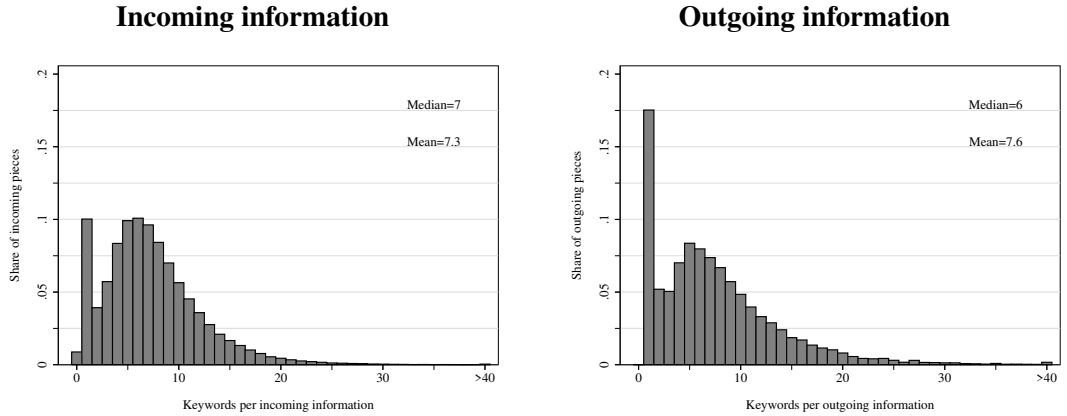
Name	# Pieces	Relevance	Most frequent institutional reference[1]	Active [Years]		
				First	Last	Period
MFAA	5944	2.5	SPD, EG, NATO	’81	’87	6
MAH	1612	2.6	EG, IWF, COCOM	’81	’87	6
AXEN	1505	2.2	NATO, SPD, EG	’81	’87	6
FISCHER	1502	2.2	NATO, SPD, EG	’81	’87	6
SIEBER	1026	2.1	NATO, PLO, EG	’81	’87	6
HONECKER	811	2.1	NATO, SPD, CDU	’81	’87	6
IPW	803	2.5	CDU, EG, BDI	’81	’87	6
ZK	763	2.3	EG, UNO, PLO	’81	’87	6
BEIL	738	2.4	EG, COCOM, RGW	’81	’87	6
KRENZ	615	2.1	NATO, SPD, WV	’82	’87	5
MITTAG	610	2.2	NATO, EG, COCOM	’81	’87	6
NVA	579	2.1	NATO, WV, BUNDESWEHR	’81	’87	6
NIER	563	2.2	NATO, SPD, CDU	’81	’87	6
SOELLE	453	2.5	COCOM, EG, RGW	’81	’87	6
STOPH	440	2.2	NATO, EG, SPD	’81	’87	6
HERRMANN	391	2.4	SPD, CDU, NATO	’81	’87	6
STRELETZ	378	1.6	NATO, WV, BUNDESWEHR	’81	’87	6
HOFFMANN	372	1.9	NATO, WV, EG	’81	’86	5
SCHALCK	304	2.2	COCOM, EG, CDU	’81	’87	6
HAGER	277	2.3	SPD, NATO, CDU	’81	’87	6

Notes: This table shows the 20 recipients who received most outgoing information between 1969 and 1987. Relevance is measured by the average relevance score of the input pieces used per outgoing piece. The HVA rated relevance on an ordinal scale with values 1 (*Very valuable*), 2 (*Valuable*), 3 (*Medium value*), 4 (*Low value*) and 5 (*Without value*). Additionally, the three most occurring institutional references among all pieces received by a recipient are given. Based on outgoing information only. **1:** See Table 3.A.4 in the appendix for further details on the listed institutions.

tion, respectively. The distribution of keywords over incoming pieces has two peaks, 1 and 6, each accounting for roughly 10% of the total, whereas, with 18% of all pieces, outgoing information is most often described by only one keyword. Figure 3.3.8 shows histograms of the respective distributions.

On average, a keyword is used 1,063 times for labelling incoming pieces and 165 times to describe outgoing information. However, the median keyword is used only 85 and 24 times, respectively. So, especially for incoming information, there is a small set of keywords which are used to label many pieces whereas the majority of distinct keywords is only used with relatively low frequency. For both, incoming and outgoing information, many keywords are only used once (4% and 10% of distinct keywords, respectively) or twice (5% and 6%, respectively). The most frequent keyword for both types of information, *Object*, is used to label 53% and 47% of pieces, respectively. Figure 3.3.9 plots the corresponding distributions of the keyword frequencies. For better readability, all values greater than 100 are excluded. The left-out parts of the distribu-

Figure 3.3.8: Empirical distribution of the number of keywords over pieces of information



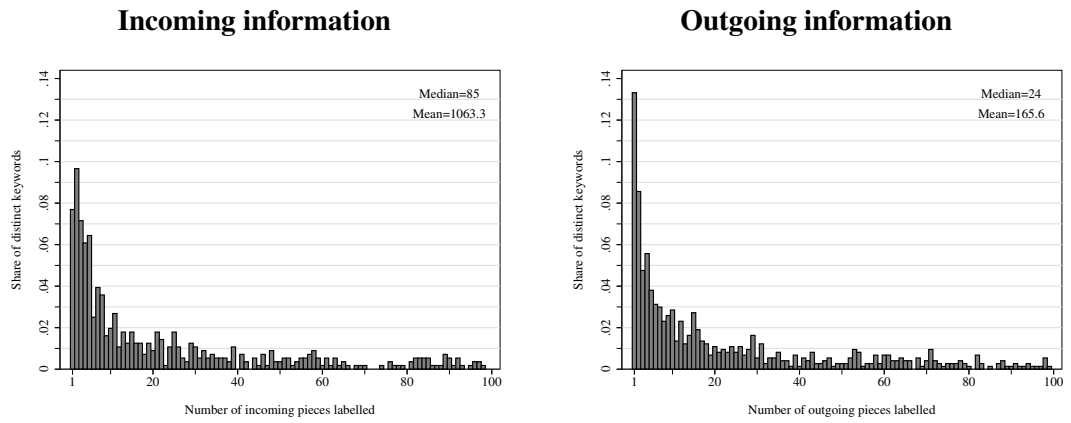
Notes: This figure shows the distribution of the number of keywords over pieces of information. For better readability, all instances of information with 40 or more keywords are shown in the rightmost bin.

tions are very spread out: the average share of keywords occurring more than 100 times is roughly 0.001 for both information types. Figure 3.A.7 in the appendix shows the same plots, however including all values greater or equal than 100 in one bin such that the relative proportions of keyword frequencies are visible.

Table 3.3.4 zooms into the right end of the distributions and shows the 30 most frequently used keywords among incoming and outgoing pieces of information, respectively, along with their English translations. Some descriptors like *Foreign policy* or *Military policy* have a relatively clear meaning. Others, like *Object*, *Name*, *N* or *SL* are more difficult to associate with sensible content. For the keyword *Object*, we find that it was very likely used to label those pieces of information for which the meta feature on institutions and organizations is non-missing.[21] Likely, the same holds true for the keyword *Name* which could have been used to indicate whether the value of the meta feature storing the names of individuals appearing in pieces of information was non-missing. As mentioned before, however, due to data protection reasons, we do not have access to this meta feature as it potentially stores clear names of IMs or their (involuntary) sources.

[21] In the SIRA terminology this meta feature is called *Object Reference*. Among all pieces of information, 4.1% are not labelled with the keyword *Object* but the value of the variable *Object Reference* is non-missing, 0.5% are labelled but the value is missing and for 95.3% the absence (presence) of the keyword indicates whether the value of *Object Reference* is missing (or not).

Figure 3.3.9: Empirical distribution of the frequency of keywords in the data



Notes: This figure shows the distribution of keyword frequencies for incoming and outgoing pieces of information. For better readability, all instances of keywords occurring more often than 100 times are excluded. The average density of these keywords is 0.001 for both plots.

3.4. Case Studies: German-German Negotiations

In this section, we show how to use the SIRA meta data to conduct qualitative research on the HVA's espionage activities. We provide evidence that, on top of spying on West German parliaments and parties, the Stasi also targeted negotiations between high-ranking East and West German representatives. As case studies, we chose the negotiations on a cultural agreement and on a loan issued to the GDR and backed by West Germany.

3.4.1. Culture negotiations

The negotiations between East and West Germany regarding a treaty on cultural exchange commenced in November 1973. These negotiations were part of a broader effort to improve the inner-German relations in various fields, a goal stipulated in the so-called Basic Treaty, signed by the GDR and West Germany in December 1972.[22] In particular, the cultural agreement was intended to facilitate, for example, the organization of art exhibitions in both parts of Germany or the common participation in international culture events.

Until 1975, delegations of both states, led by the West German ambassador in the GDR, Günter Gaus, and the deputy minister for foreign affairs of the GDR, Kurt Nier,

[22]The Treaty concerning the basis of relations between the Federal Republic of Germany and the German Democratic Republic (*Grundlagenvertrag* or *Vertrag über die Grundlagen der Beziehungen zwischen der Bundesrepublik Deutschland und der Deutschen Demokratischen Republik*) marked the beginning of official diplomatic relations between both German states. Next to collaboration in cultural matters, it was also agreed to start negotiations regarding economic exchange, traffic, transit and postal services.

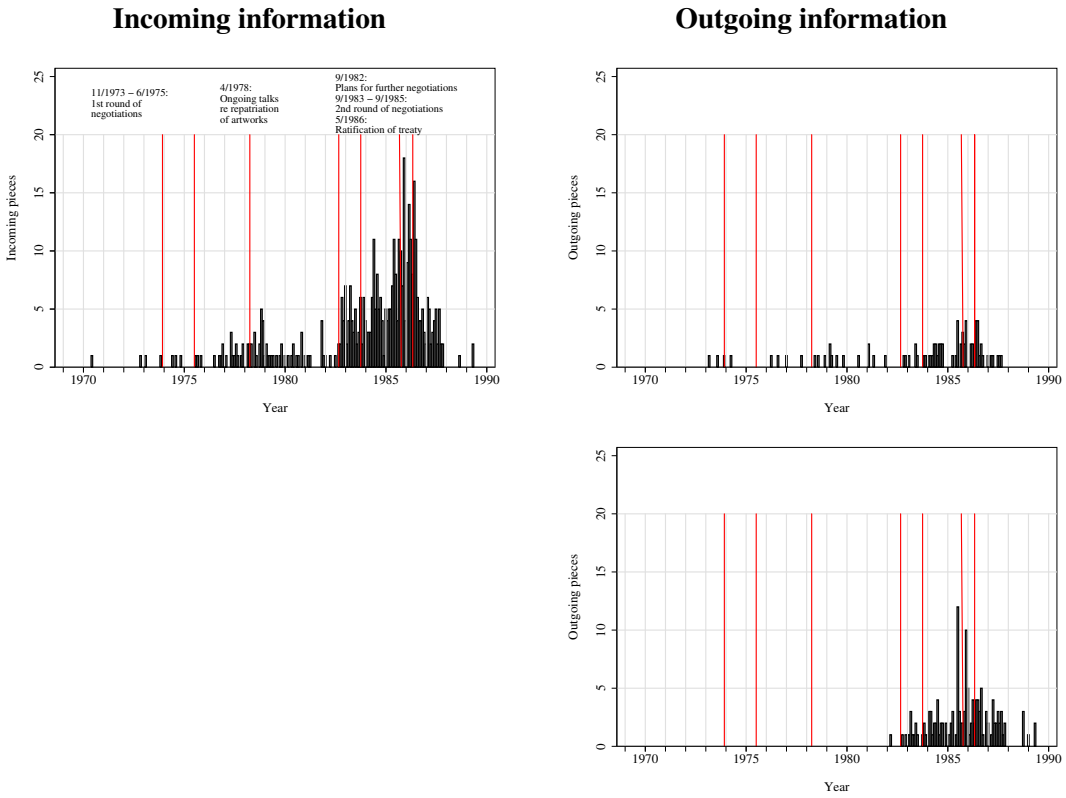
Table 3.3.4: Most frequent keywords

Keyword		Translation	% of pieces labelled		
			Incoming	Outgoing	(rank)
1	OBJEKT	Object	53.0	47.0	(1)
2	NAME	Name	28.0	15.0	(5)
3	AUSZENPOLITIK	Foreign policy	24.0	19.0	(3)
4	INNENPOLITIK	Domestic policy	15.0	10.0	(16)
5	EINSCHAETZUNG	Assessment	15.0	18.0	(4)
6	PARTEI	Party	14.0	10.0	(15)
7	BEZIEHUNG	Relation	13.0	7.0	(22)
8	WIRTSCHAFT	Economy	12.0	11.0	(14)
9	ZUSAMMENARBEIT	Collaboration	12.0	11.0	(13)
10	DIFFERENZEN	Differences	12.0	13.0	(8)
11	MILITAERWESEN	Military affairs	12.0	13.0	(9)
12	FEINDTAETIGKEIT	Enemy activity	11.0	12.0	(11)
13	HALTUNG	Stance	11.0	9.0	(18)
14	OST-WEST-BEZIEHUNG	East-West-Relation	11.0	14.0	(6)
15	STAATSAPPARAT	State apparatus	11.0	14.0	(7)
16	STREITKRAEFTE	Armed forces	10.0	12.0	(12)
17	REGIERUNG	Government	10.0	13.0	(10)
18	N	N	9.0	19.0	(2)
19	HANDEL	Trade	9.0	10.0	(17)
20	WIRTSCHAFTSPOLITIK	Economic policy	8.0	8.0	(21)
21	SICHERHEITSPOLITIK	Security politics	8.0	7.0	(23)
22	INNERE LAGE	Interior affairs	8.0	8.0	(20)
23	KONFERENZ	Conference	7.0	6.0	(26)
24	WAHL	Election	7.0	4.0	(34)
25	MILITAERPOLITIK	Military policy	7.0	6.0	(29)
26	RUESTUNG	Armament	7.0	7.0	(25)
27	BEZIEHUNGEN	Relations	6.0	9.0	(19)
28	ABRUESTUNG	Disarmament	6.0	5.0	(32)
29	INDUSTRIE	Industry	6.0	6.0	(28)
30	MILITAERTECHNIK	Military technology	6.0	7.0	(24)
33	SL	SL	5.0	6.0	(27)
40	FUEHRUNG	Command	4.0	5.0	(30)

Notes: This table shows the percentage of pieces of information labelled by the 30 most frequently used keywords for incoming pieces. The percentage and rank (in parentheses) of the same keywords among outgoing pieces of information is also given. The last two keywords are shown because they are among the 30 most frequent keywords among outgoing pieces of information.

met five times. These discussions were not very fruitful, especially since a central demand of the GDR - the repatriation of certain artworks that had been relocated to West Germany during and immediately after the second world war - was not met by West German representatives. In September 1982, however, the GDR agreed to proceed with the negotiations explicitly excluding the repatriation topic. Now led by the new West German ambassador in the GDR, Hans Otto Bräutigam, the West and East German delegations (the latter still led by Kurt Nier) met for another 12 negotiation rounds during the

Figure 3.4.1: Pieces of information related to culture negotiations



Notes: The upper plots show the number of incoming and outgoing pieces of information labelled with keywords (*Culture* or *Cultural policy*) and (*Treaty* or *Negotiation* or *Negotiations*). The bottom plot shows the number of outgoing pieces of information labelled with keywords (*Culture* or *Cultural policy*) and sent to Kurt Nier (lead negotiator GDR), Karl Seidel (high-ranking GDR diplomat and regular participant of the negotiations) or to the MFAA (East German ministry for foreign affairs).

years 1982 to 1985. In May 1986, both parties signed the treaty on cultural exchange.[23]

To find the information relating to these negotiation series, we first filter all pieces by keywords keeping only those labelled with *Culture* or *Cultural Policy* or both. From this sample, we discard all pieces that are not labelled either with keywords *Treaty*, *Negotiation* or *Negotiations*. The upper plots in Figure 3.4.1 plot the monthly distribution of the remaining pieces of information together with some of the previously mentioned dates of the culture negotiations.

While the number of relevant pieces of information was relatively low during the first five negotiation rounds between 1973 and 1975, it surged right after the GDR signalled its willingness to continue the negotiations in September 1982. Once the treaty was signed in 1986, the number of pieces of information declines again. Unfortunately, we are unable to explain the inflow and outflow of information with the relevant keywords between 1976 and 1982, peaking in 1979. Likely, these pieces were produced in

[23]The preceding paragraph is based on Lindner (2011) and Bracher and Jacobsen (2014).

the context of the ongoing dispute on the repatriation of displaced artworks.

The lower right panel of Figure 3.4.1 plots the monthly distribution of outgoing pieces of information that are labelled with keywords *Culture* or *Cultural policy* and for which the recipient is either Kurt Nier, Karl Seidel (high-ranking GDR diplomat and regular participant of the negotiations) or the MFAA (East German foreign ministry, *Ministerium für Auswärtige Angelegenheiten*). The pattern is very similar to that of the upper right plot, however there are no pieces of information recorded before 1982. This might be a hint that the information picked up in Figure 3.4.1 between the end of the first negotiation rounds and 1982 is not related to the culture negotiations. On the other hand, data on recipients of information is available only for roughly 25% of outgoing pieces registered before 1983 so that the pattern might result from missing data.

3.4.2. West German loan to GDR

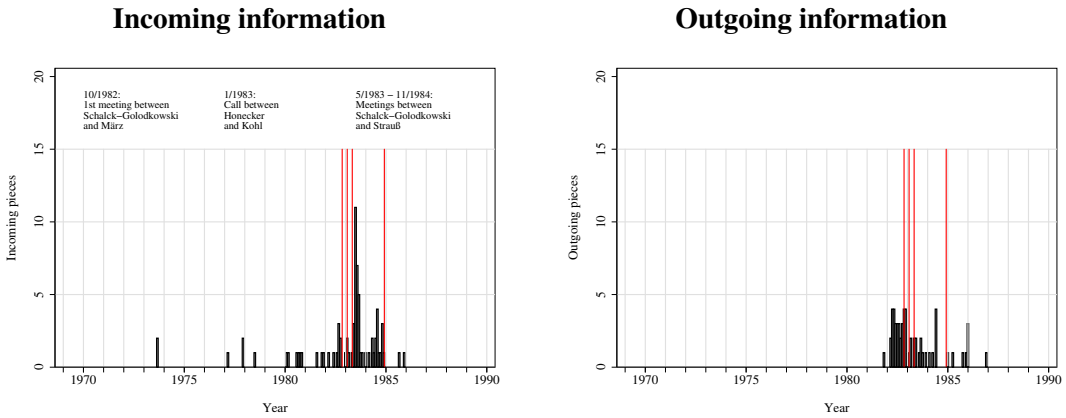
In 1983, the countries of the Eastern Bloc experienced a severe economic crisis which had already forced Poland to declare national bankruptcy. To avoid a similar fate, SED leader Erich Honecker mandated MFAA division head Alexander Schalck-Golodkowski to negotiate terms for a loan backed by West Germany with Bavarian prime minister Franz Josef Strauß. The negotiations were kept secret, to the extent that even West German foreign and finance ministers Genscher and Stoltenberg were not informed by chancellor Kohl. The contact between Schalck-Golodkowski and Strauß was facilitated by meat wholesaler Josef März who knew Schalck-Golodkowski through his business in East Germany and was also a friend of Strauß. März and Schalck-Golodkowski had a series of four meetings beginning in October 1982 regarding the loan issuance. In January 1983, part of the negotiations were conducted on the highest political level when Honecker and Kohl exchanged their respective positions on the matter during a phone call. After further preparatory meetings, Schalck-Golodkowski, Strauß and Philipp Jenninger (State Secretary of chancellor Kohl) finalized the terms of the loan contract during three meetins in May and June 1983.

On 1 July 1983 it was publicly announced that a consortium of West German banks would issue a loan worth 1 billion Deutsche Mark backed by West Germany to the GDR. From Septermber 1983 to November 1984, Schalck-Golodkowski and Strauß continued their meetings to negotiate and discuss another loan which was issued in July 1984.[24]

Figure 3.4.2 plots the monthly distributions of incoming and outgoing pieces of information, respectively, which we deem related to the previously described negotiations. For incoming information, we keep only those pieces labelled with keywords *Loan*,

[24]The preceding paragraph is based on BStU (2010), Eisenbichler (2012), Bracher and Jacobsen (2014) and Oetzinger (2016).

Figure 3.4.2: Pieces of information related to loan negotiations



Notes: This figure shows the number of incoming and outgoing pieces of information labelled with keywords *Loan*, *Loans* or *Issuance of loan*. Additionally, the left panel shows only incoming pieces of information associated with the West German CSU and the right panel only outgoing pieces of information sent to Schalck-Golodkowski (lead negotiatior GDR) or Honecker.

Loans or *Issuance of loan*. We suspect that pieces directly related to the negotiations between Schalck-Golodkowski and Strauß were likely to contain some information on the West German negotiators. Since we do not observe the *Name* meta feature, we instead consider only pieces associated with the CSU, the party led by Strauß. While some information with these characteristics already arrived at the Stasi as early as 1973, one can see that deliveries became more frequent in 1980 and significantly increased during the period in which Schalck-Golodkowski and Strauß were negotiating.

For outgoing pieces, we filter for the same keywords, however, instead of requiring pieces to be associated with the CSU, we keep only those that were addressed to Schalck-Golodkowski. The outgoing pieces satisfying our criteria were exclusively created in the period around the credit negotiations.

3.5. Topic model

In this section, we show how to use the keywords with which pieces of information were labelled to perform a quantitative analyses of the thematic aspects of the Stasi's espionage activities. To do so, we perform a topic analysis with Latent Dirichlet Allocation (LDA). LDA is a type of latent semantic analysis (Deerwester et al., 1990) which is used to retrieve information from textual data. It was developed by Blei, Ng, and Jordan (2003) as an extension to the work of Hofmann (1999) who introduced probabilistic theory to latent semantic analysis. Although there have been advances in probabilistic latent semantic analysis, LDA still remains state of the art in topic modelling (Lüdering

and Tillmann, 2020).

3.5.1. Estimation

The basic assumption of LDA is that the content of text documents, or just *documents*, is comprised of several unobserved, that is latent, *topics*. In the context of LDA, a topic is defined as a probability distribution over all distinct words, the *vocabulary*, which occur in a collection of documents, the text *corpus*. Within the SIRA database, each piece of information constitutes a document, the distinct keywords are the vocabulary and together the collection of pieces of information labelled with keywords form the corpus.

To understand the concept of a topic, assume for the moment that Stasi espionage was only focused on two aspects: military and political activity in West Germany. An appropriate LDA model would identify those two topics by generating two probability distributions over the entire set of keywords: one distribution would feature a lot of probability mass on military-related keywords like *Military Technology* or *Military Affairs* but only little mass on unrelated terms like, for example, *Election*. The second distribution would assign probabilities in the opposite way. When presented with these probability distributions, the researcher must infer or label the underlying topic herself. The LDA model also provides, for each document, a probability distribution over the found topics. For example, a piece of information labelled exclusively with keywords *Military technology* and *Military affairs* will likely exhibit a high probability mass on the topic related to military activity but a low or zero mass for the topic capturing political activity.

In order to produce these results, LDA assumes an imaginary random process which generates the text corpus. Each document in the corpus consists of a mixture of a fixed set of different topics. Each word within a document is generated by drawing a topic, given a per-document distribution over topics, and then drawing the word itself from the chosen topic, which, recall, is itself a distribution over words. Further, LDA assumes that each of these latent, i.e. unknown, distributions is a Dirichlet. This generative process implies a joint probability distribution of observed (the corpus) and unobserved random variables. The goal is to infer the conditional distribution of the hidden variables (the parameters of the Dirichlets), given the observed textual data and, obviously, the assumptions on the generative process. The result are several distributions over the vocabulary, the topics, which (approximately) maximize the probability to observe the corpus. Intuitively, very similar to clustering algorithms, LDA infers topics from words which frequently occur together in the same documents. However, standard maximum likelihood techniques cannot be applied because the model is too complex. Instead, in

recent applications, the likelihood function is approximated via Gibbs sampling, a type of Markov Chain Monte Carlo approach (Griffiths and Steyvers, 2004).[25]

Often, LDA is applied to corpora with documents comprised of many words like scientific papers (Blei, Ng, and Jordan, 2003; Griffiths and Steyvers, 2004). Our text data is significantly different because an average document, i.e. piece of information, exhibits only roughly 7 keywords. After data pre-processing (described later in this section), this number goes up to 8.4. In that, keywords associated with pieces of information are more similar to Twitter posts (*Tweets*) which feature an average length of roughly 12 to 14 words (Boot et al., 2019). This similarity is confirmed by Meyer et al. (2019). In their paper, the number of distinct words equals 5,076 and the number of Tweets is 87,030. According to the authors, for longer texts, usually, the number of distinct words is much larger than the number of documents. In our data, after pre-processing, we are left with a corpus of 1,007 distinct keywords and 104,442 pieces of information. While the implied ratio of these values in our data (0.01) is roughly 6 times smaller than that found by Meyer et al. (2019), our corpus still exhibits the key feature that the vocabulary size is considerably smaller than the number of documents. Due to these similarities, we will adjust our implementation of the LDA model accordingly.[26]

Further, note that, in order to consistently compare topics across incoming and outgoing information, we estimate the model based on keywords appearing in both types of pieces of information. We use Python's *sklearn* library to implement our LDA model.

Data pre-processing Usually, for semantic analysis of textual data with LDA, a lot of pre-processing is necessary. For example, *stemming* is performed so that different grammatical forms of the same word are only counted once. Given our standardized data, this is not an issue, however, we manually adjust some keywords: for example, as shown in Table 3.3.4, *Relation* and *Relations* are two distinct keywords. Here, and for all other similar cases, we transform the plural form to its singular version. We end up relabelling 51 keywords manually. Table 3.B.1 in the appendix shows the exact changes performed. Additionally, we remove all empty spaces between keywords so that the algorithm recognizes them as one word. In our context, this is very easy because the data is structured such that we can identify words which, together, form one keyword. For example, we know that the three words *Asian Security System* are forming only one keyword so that we can pass them to the algorithm as *AsianSecuritySystem*. In other applications, word pairs which form one semantic unit, so-called bi-grams, are

[25] The preceding paragraphs are based on Blei, Ng, and Jordan (2003) and Lüdering and Tillmann (2020).

[26] All of these adjustments will be explicitly mentioned. Some of these adjustments are only mentioned in the model selection section in the appendix.

more difficult to identify. For convenience, we also remove any punctuation, hyphen or underlines. After these modifications, we are left with 178,131 (incoming and outgoing) pieces of information, labelled by 1,300,904 total and 1,033 distinct keywords.

The fact that the labelling of outgoing information differs from that of incoming pieces of information (see Figures 3.3.8 and 3.3.9) might be a sign that spies did not describe their gathered information with keywords as carefully as the Stasi did with the outgoing pieces. This might distort our topic analysis. Therefore, we use the meta feature which classifies the reliability of spies according to 5 categories. We drop incoming pieces of information for which we cannot observe the respective spy or her reliability score as well as those pieces of information gathered by spies graded with either of the lowest three reliability classes. From the 178,131 pieces of information, 38,790 are removed in this step. This corresponds to 210,599 keywords, however the number of distinct keywords remains at 1,033.

Very frequently occurring words like prepositions or auxiliary verbs which do not add to the specific meaning of a topic are usually removed in topic modelling. Since our textual data does not contain such so-called *stop-words*, this step is unnecessary. However, some keywords still appear so often and add so little meaning that they might affect the accuracy of LDA. Keywords *Object* or *Name* are obvious examples. A similar problem arises for words occurring only very infrequently, which many words in our data do (compare Figure 3.3.9). We follow Blei and Lafferty (2009) and Hansen, McMahon, and Prat (2018) in ranking all keywords using term frequency-inverse document frequency (tf-idf), a score measuring information content and punishing too frequent and infrequent keywords. We drop all those keywords with a score lower or equal to the 5% percentile. In doing so, we remove 26 distinct and 148,520 total keywords. Since there are pieces of information which are labelled exclusively by keywords with too low tf-idf scores, we also lose 656 pieces of information.

Since LDA clusters words based on joint occurrences in documents, we also drop all pieces of information described by only one keyword which survived the previous step. Note that, in this step, we might also drop pieces of information which originally were associated with more than one keyword but, after the previous step, were left with only one keyword describing them. Tang et al. (2014) mention that document length should be at least on the order of the logarithm of the number of documents. After the previous steps we are left with 138,685 pieces of information, i.e. documents, the logarithm of which equals roughly 12. In order not to drop too many documents (compare Figure 3.3.9), we decide to exclude those pieces of information labelled by 3 or less keywords.

After these adjustments, we are left with a corpus of 104,442 pieces of information, 878,560 total and 1,007 distinct keywords. On average, each piece of information is

labelled by 8.4 keywords, the maximum and minimum number of keywords per information is 4 and 68, respectively.

Model selection The maximization of the likelihood function implied by the LDA model requires an a priori selection of three different so-called *hyper parameters*. The number of topics, i.e. clusters, the model should produce and two parameters determining the sparsity of the involved Dirichlet distributions. When the Dirichlet distributions are assumed to be symmetric, which is the case here, the first parameter, α, determines how many topics are likely to occur in each document. A high value implies that each document is a mixture of many topics, the distribution is not very sparse. If the value is low, only some topics are likely to be present per document, a sparse distribution. The second parameter η (sometimes also β), regulates the mixture of words per topic: a higher value implies that a topic is comprised by many words whereas a low value results in fewer important words per topic.

To select the appropriate combinations of parameters, we follow a very common approach in the literature and cross-validate the predictive power of the LDA model on a held-out dataset with the help of a measure called Perplexity (see, for example, Manning and Schutze, 1999, Griffiths and Steyvers, 2004, Hoffman, Bach, and Blei, 2010, Tang et al., 2014, Hansen, McMahon, and Prat, 2018). Perplexity is simply a function of the likelihood to observe the held-out data, given the previously specified LDA model. We also consider the interpretability of our results. Appendix 3.B.1 contains a detailed description of our cross-validation approach and the respective references. In the end, we opt for a model using 11 different topics and setting $\alpha = 0.031$ and $\eta = 0.033$.

3.5.2. Outcome

Recall that the outcome of the topic model, 11 topics, are 11 different distributions over the entire vocabulary, i.e. those keywords not dropped during pre-processing. LDA generates these distributions based on the co-occurrence of keywords in pieces of information. Table 3.5.1 presents, for each of the topics, the 15 keywords associated with the highest probability of occurring within a topic.

First, we label the topic identified by keywords *State apparatus*, *Government*, *Relation*, *Assessment*, *Stance*, *East-West relation*, *Domestic policy* and *Journey* as *Relation to West German government*. Pieces labelled by a combination of these keywords likely contain information on the stance of the West German government regarding different policy areas. The descriptor *Journey* probably refers to diplomatic visits of East and West German government officials in the respective other country.

The second topic describes the *Political landscape in West Germany*. Keywords

Table 3.5.1: Top words per topic, 11 topics

1: Relation to West German government	**2: Political landscape in West Germany**	**3: Geo & security politics**	**4: Trade & Economic relation with West Germany**
State apparatus	Party	Security politics	Trade
Government	Election	East-West relation	Economy
Assessment	Domestic policy	Disarmament	Collaboration
Relation	State of interior affairs	Armament	Economic policy
Domestic policy	Differences	Stance	Relation
Differences	Parliament	Negotiation	Development policy
Stance	Assessment	Detente	Assessment
Journey	KP	KSZE	Payment transactions
Party	SP	Conference	Financial concerns
Coalition	Election campaign	SL	Industry
East-West relation	Coalition	Assessment	East-West relation
State of interior affairs	Party congress	MVM	Export
Election	Command	Party	Loan
Conception	Organisation	Nuclear weapon	Agriculture
Economic policy	Union	Check	Investment
5: Civil R&D (energy)	**6: Civil R&D & military technology**	**7: Military strategy & technology**	**8: Military strategy & international conflicts**
Commodity	Research	Armed forces	Relation
Energy	Armament	Military affairs	Differences
Nuclear energy	Development	Military technology	Military affairs
Economic policy	Industry	Military policy	Stance
Industry	Electronics	Land forces	Conflict settlement
Energy policy	Collaboration	Air force	Middle-East policy
Crude oil	Firm	Armament	Domestic policy
Energy economy	Military technology	Assessment	Armed forces
Economy	Project	Planning	Military policy
Assessment	IT	Reconnaissance	State of interior affairs
Research	Financing	Naval forces	Liberation movement
Collaboration	Economy	Collaboration	Conflict
Environmental protection	Science	Military strategy	Assessment
Nuclear research	Electrotechnology	ELoKa	War
Nuclear industry	Report	Missile	Arms supply
9: Enemy activity (foreign nations (& intelligence)	**10: Enemy activity civil society)**	**11: Enemy activity (Media)**	
Enemy activity	Enemy activity	Enemy activity	
Security	Domestic policy	PID	
Enemy organisation	Difficulty	Relation	
Intelligence agency	Relation	Contact	
Counter revolution	State of affairs	Media	
Military affairs	Assessment	Information technology	
Law	SL	West Berlin status	
Terrorism	Population	Transportation	
State apparatus	Economy	Organisation	
Domestic policy	PID	Treaty	
Measure	Ostforschung	Ostpolitik	
State of interior affairs	Church	Stance	
Main problem	East-West relation	Socialist countries	
Extremism	Situation	Press	
ZV	Youth	Assessment	

Notes: This table shows the English translation of the 15 most important keywords for each of 11 topics identified by an LDA model with $\alpha = 0.03125$ and $\eta = 0.0328125$.

Party, *Election*, *Domestic policy*, *State of interior affairs*, *Differences*, *Parliament*, *Election campaign*, *Coalition* or *Party congress* appear within the list of top 15 words. Pieces of information labelled with a combination of these descriptors likely contained details

on West German parliaments and parties, and their relationship with each other. The Stasi's espionage on the West German federal parliament in Bonn, uncovered by BStU (2013a), is probably also attributable to this topic.

Third, we find that keywords *Security policy*, *East-West relation*, *Disarmament*, *Armament*, *Stance*, *Negotiation*, *Detente*, *KSZE*, *Conference* or *MVM* form a stable topic which we label *Geo & security politics*. *KSZE* is the German abbreviation for the Conference on Security and Co-operation in Europe (CSCE), a series of summits between the NATO and the states of the Warsaw Pact which was integral to the detente process during the Cold War. Keyword *MVM*, the abbreviation for *Militärverbindungsmission*, Military liaison missions in English, describes regular military missions which were intended to allow representatives of Western and Soviet military intelligence to monitor each other's occupied parts of Germany.

We label the fourth topic as *Trade & Economic relation with West Germany*. Notably, its most defining keywords appear across two topics when allowing a higher number of total topics. In Appendix 3.B.1 we find that keywords *Trade*, *Export*, *Import*, *(East-West) relation* and *Commodity* are characteristic for a topic we call *Trade with West Germany*. Keywords *Economy*, *Economic policy*, *Collaboration*, *Relation*, *Financial affairs*, *Development policy*, *Industry*, *Conference*, *Investment*, *Development aid* and *Entrepreneur's association* define a topic we call *Economic relation with West Germany*. Clearly, the topic *Trade & Economic relation with West Germany* represents a conjunction of these two topics.

The fifth topic identified by the LDA model could be described as *Civil R&D (energy)*. It is characterized by keywords *Commodity*, *Energy*, *Nuclear energy*, *Industry*, *Energy policy*, *Energy economy*, *Nuclear research*, *Nuclear industry* and *Crude oil*.

Topics six, seven and eight, again, represent conjunctions of topics that appear as stand-alone for models with a higher number of total topics. Topic six is a mixture of keywords characteristic for topics *Civil R&D* and *Military technology*, respectively. The former consists of descriptors *Science*, *Research*, *Development* or *Firm*. The latter uses keywords like *Military technology*, *Research*, *Development*, *Armament*, *Project* and more concrete keywords like *Missile*, *Airplane* or *Weapon*.

Topic seven captures pieces of information that were related to military technology and a further topic we call *Military strategy*. The latter is identified by keywords *Military affairs*, *Military policy*, *Land forces*, *Military strategy*, *Air force*, *Exercise*, *Command*, *Defense* and *Naval forces*. We decide to label the conjunction of those two topics as *Military strategy & technology*

Topic eight adds to the keywords of the topic *Military strategy* those that can be associated with international conflicts. These are *Relation*, *Differences*, *Conflict settle-*

ment, *Stance*, *Domestic policy*, *Middle-East policy*, *Conflict*, *War*, *Terrorism*, *Liberation movement*. Accordingly, we label the respective joint topic as *Military strategy & international conflicts*

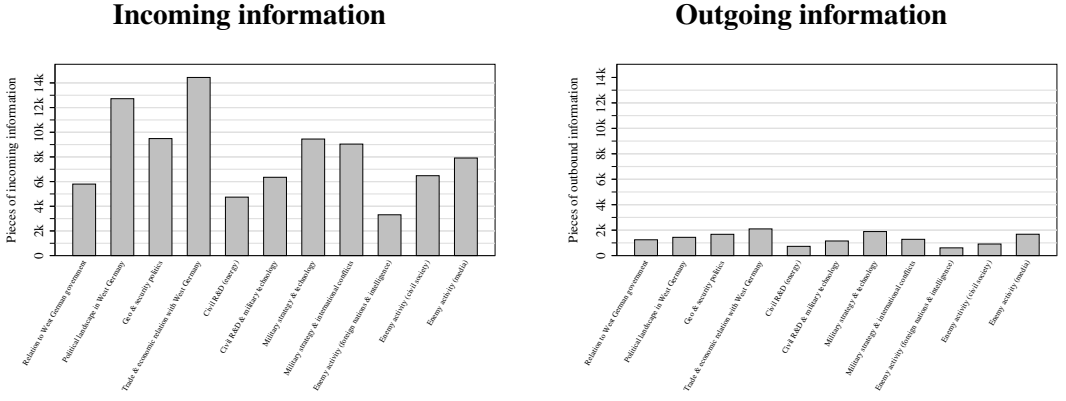
Finally, topics nine to eleven all feature the keyword *Enemy activity*, *Feindtätigkeit* in German, an established term within the MfS (Müller-Enbergs, n.d.). Topics ten and eleven further exhibit the descriptor *PID* as an important keyword. PID is an abbreviation for *Politisch-ideologische-diversion* (Political-ideological-diversion), also a very common part of the Stasi terminology, which was subsequently adopted by many communist security services (Engelmann, n.d.). The term was used to label the ideological influence of the so-called West and smaller entities (e.g. institutions or individuals whose ideology deviated from the SED's official doctrine) on communist societies. When the above mentioned keywords are accompanied by descriptors *Enemy organisation*, *Secret service* or *State apparatus*, we label the associated topic as *Enemy activity (foreign nations & intelligence)*. Topic ten, additionally featuring descriptors *Population*, *Church* or *Youth*, we label as *Enemy activity (civil society)*. Topic eleven, including descriptors indicating the involvement of the media, e.g. *Media*, *Press* and *Information technology*, we name *Enemy activity (media)*.

Appendix 3.B.1 additionally presents results for topic models with a total of 10 to 15 topics. We also provide a comprehensive description, especially on how single topics change when increasing the total number of topics.

Number of pieces per topic To gain an overview of the focus of the Stasi's political espionage activities, we use the estimated LDA model to associate each piece of information that was part of the input data with the topics it is likely to cover. The respective output of the model is a probability distribution over each topic for each piece of information. Figure 3.5.1 shows the result of aggregating the per-information probability distributions. The vertical axis of the respective bar graphs can be interpreted as the number of pieces of information dealing with each of the previously described topics.

The ranking in terms of numbers of pieces by topic is relatively similar between incoming and outgoing information. With more than 12,700 and 14,400 incoming pieces, respectively, most information was either related to the *Political landscape in West Germany* or covered content on the *Trade & economic relationship with West Germany*. Topics *Geo & security politics*, *Military strategy & technology*, *Military strategy & international conflicts* and *Enemy activity (media)* follow with a total of more than 9,400, roughly 9,400 and roughly 7,900 pieces. About 6,400, 6,300 and 5,800 pieces each, were dedicated to *Enemy activity (civil society)*, *Civil R&D & military technology* and *Relationship with West German government*. The least occurring topics, with roughly

Figure 3.5.1: Pieces of information by topic



Notes: This figure shows the number of incoming and outgoing pieces of information associated with each topic identified by the LDA model.

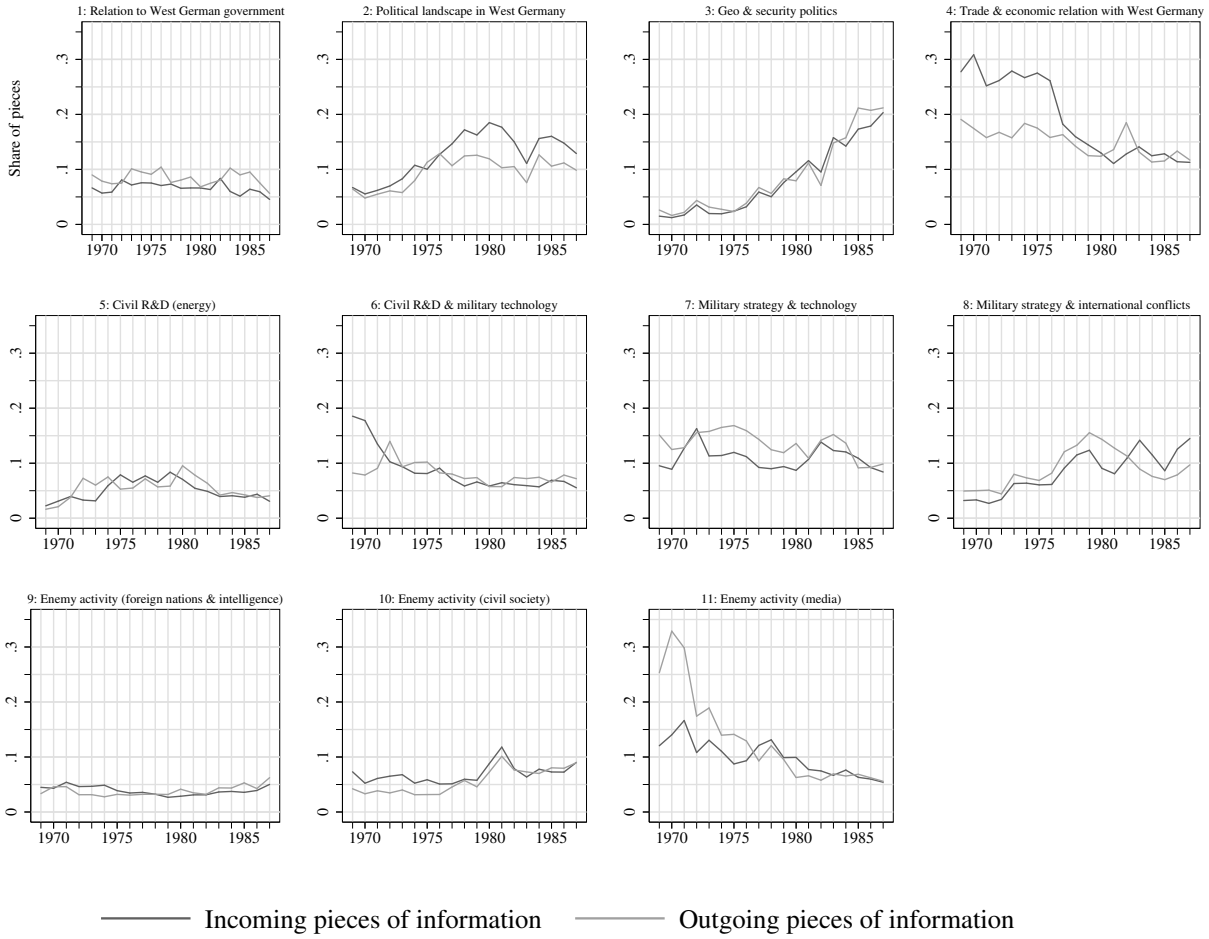
5,000 and 3,000 pieces, respectively, were associated with *Civil R&D (energy)* and *Enemy activity (foreign nations & intelligence)*. Note, however, that together the pieces of information related to *Enemy activity* make up the biggest share among all incoming pieces. The most striking difference in the topic ranking between incoming and outgoing information occur for topics *Political landscape in West Germany* and *Geo & security politics*: while the former makes up a larger share of incoming pieces than the latter, this relationship is reversed for outgoing pieces.

To see the development over time, Figure 3.5.2 plots, for each topic, its share in the total number of pieces of information across years. While topics *Relation to West German government*, *Civil R&D (energy)*, *Military strategy & technology*, *Enemy activity (foreign nations & intelligence)* and *Enemy activity (civil society)* maintain a relatively constant share of total incoming and outgoing pieces of information, there are significant trends for other topics. Pieces of information dealing with the *Political landscape in West Germany* accounted four roughly 5% of incoming and outgoing pieces in 1969 but for about 18% and 11%, respectively, in 1980. After a significant drop in 1983, the topic made up 14% and 10% of incoming and outgoing pieces in 1987, respectively.

Pieces on *Geo & security politics* had a share in total incoming and outgoing pieces of less than 5% in 1969, however, their number increased continuously over years such that in 1987, 20% of incoming and outgoing pieces were dealing with *Geo & security politics*.

The topic *Trade & economic relation with West Germany* accounted for roughly 28% of incoming and 20% of outgoing pieces in 1969, however after a steady decline over years, in 1987, the topic only made up slightly more than 10% of total incoming and outgoing pieces.

Figure 3.5.2: Pieces of information by topic and year



Notes: This figure shows the share of incoming and outgoing pieces of information associated with each topic in the total number of incoming and outgoing pieces of information across different years.

The share of incoming pieces associated with *Civil R&D & military technology* declined from slightly less than 20% in 1969 to about 5% in 1987. For outgoing pieces the respective share remained relatively constant around 7%.

Incoming pieces of information associated to the topic *Military strategy & international conflicts* made up roughly 4% of total incoming pieces in 1969 and increased this share to about 15% in 1987. For outgoing pieces the increase in the same period is slightly less pronounced, starting at 5% and ending at 10%, however, there is a peak in 1979 when 15% of all outgoing pieces were associated with the topic *Military strategy & international conflicts*.

Finally, the topic *Enemy activity (media)* accounted for 12% and 32% of all incoming and outgoing pieces, respectively, in 1969, however that share dropped to only 5% in 1987.

Figures 3.B.2 and 3.B.3 in the appendix show similar plots for the number and share of spies collecting pieces of information associated with each topic, suggesting that a key channel through which the Stasi managed to shift its thematic priorities over time was by placing new informants in positions of interest.

Relevance of topics To additionally compare the importance of topics with respect to their relevance in the eyes of the Stasi, we compute, for each topic, its average relevance score. Figure 3.5.3 shows that almost all topics have an average score between roughly 2.6 to 2.7, corresponding to a rating of information value between *valuable* and *medium value*. The topic *Military strategy & technology*, however, exhibits an average relevance score of 2.3, indicating that associated pieces of information were of higher importance to the Stasi than those pieces referring to the other topics.

We also compute the yearly average relevance score assigned to incoming pieces by HVA evaluators. Figure 3.5.4 shows the results. The average yearly relevance scores for all topics fluctuate between 2.5 and 3, except for the topic *Military strategy & technology* where it reaches a value of 2.2 in 1977 (recall that lower values imply higher information value).

For topics 1, 2, 4, 5, and 11 we observe similar dynamic profiles: relevance scores are around 3 in 1969, increase by roughly 0.1 until 1971/1972, then plummet to values around 2.7 between 1972 and 1974, remaining relatively constant thereafter. Topics 3 and 8 exhibit similar patterns but their average relevance scores in 1969 are substantially lower with values of 2.7 and 2.8, respectively, and drop more steeply in 1972, reaching values of 2.4 and 2.5, respectively. Afterwards, both values increase to values of around 2.6.

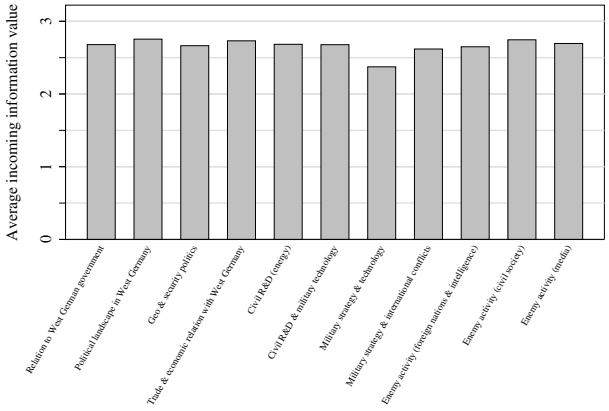
Topics 6, 7, 9 and 10 also exhibit steeply declining relevance scores in 1972 to 1973 which are followed by a one to two year period of increases and then a further relatively sharp drop in 1977 to 1979. Afterwards the topic's relevance scores stabilize.

Comparing Figures 3.5.2 and 3.5.4, we conclude that trends in the number of pieces of information collected per topic do not reflect the respective trends in average relevance scores; there does not seem to be a quantity-quality trade-off in the gathering espionage-based information.

3.6. Conclusion

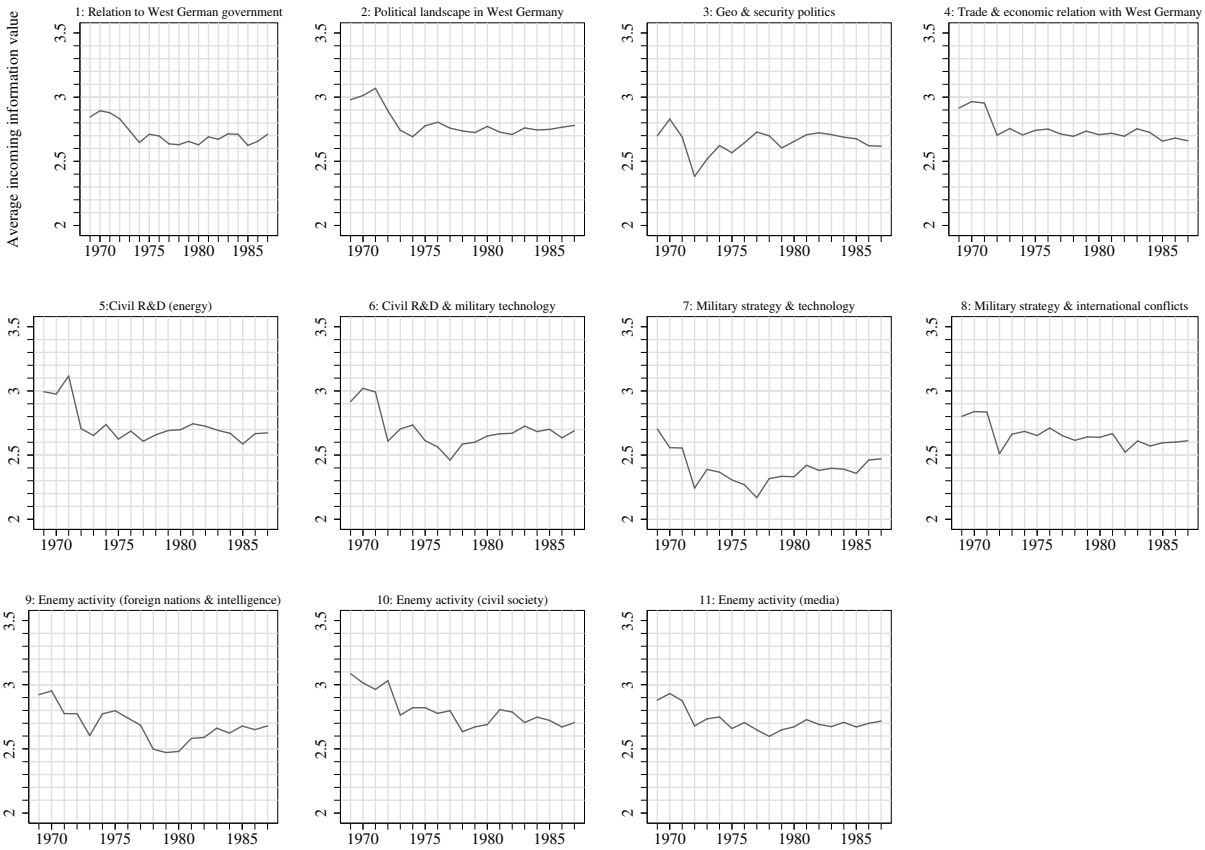
In this paper, we provide an overview of the political espionage activity of the East German secret service's foreign intelligence unit, the HVA. We show that East Germany received vast amounts of potentially relevant information through its spies located in the

Figure 3.5.3: Average relevance by topic



Notes: This figure shows the average relevance score for each topic.

Figure 3.5.4: Average relevance by topic and year



Notes: This figure shows the average relevance score for each topic across different years.

West, especially between 1977 and 1987. Most of these materials were either copies of original documents or reports written by the spies or their HVA case officers. The majority of information was related to Central Europe and West Germany and covered organizations like the NATO and leading West German parties.

We also show that the average spy in the West was active only for a period of 2 years and delivered roughly 12 pieces of information. A handful of exceptional spies, however, were active significantly longer (often more than 15 years) and provided the HVA with more than 1,000 pieces of information.

The gathered information was analyzed by dedicated HVA employees and then forwarded to its recipients, who were often SED party members or friendly intelligence agencies like the KGB. Similarly to the spies, the average recipient received only 192 pieces of information, but a small set of individuals and institutions received considerably more materials, for example the GDR ministry of foreign affairs.

With two analyses, we provide a starting point for further research efforts and show how to exploit especially the keyword data for quantitative investigations. First, we provide evidence that the HVA targeted negotiations between high-ranking East and West German officials. We filter the inflow of information by relevant keywords and recipients and show that the resulting time series of pieces of information roughly matches the negotiation dates on a cultural agreement and a significant loan issued to the GDR and backed by West Germany.

Second, we use the corpus of keywords to estimate an LDA model to uncover the thematic emphasis of the HVA's espionage activity. We find that 11 different topics fit the data best. While the topics *Political landscape in West Germany* and *Trade & economics relation with West Germany* are covered by most pieces of information, in terms of informational value, the topic *Military strategy & technology* was deemed most important by the HVA.

Future research will have to show whether the unique information of the SIRA data can be used to uncover interesting causal relationships in the realm of political espionage.

Appendix

3.A. Appendix to Section 3.3

Table 3.A.1: Overview of available meta features

Meta feature	Description	% of pieces of information with non-missing value[1]		
		Total	Inc.	Out.
1	Unique identifier (also distinguishes incoming from outgoing information)	100.0	100.0	100.0
18	**Inc.:** Date at which the arrival of incoming information was acknowledged by dedicated full-time HVA employee **Out.:** Date at which the outgoing information was finalized	99.3	99.2	100.0
33	**Out.:** HVA department responsible for creating the outgoing information	12.3	0.0	100.0
60	**Inc.:** Full-time HVA employee responsible for IM that collected the incoming information **Out.:** Full-time HVA employee(s) responsible for IM(s) that collected the incoming information on which the outgoing information was based on	90.4	99.4	26.4
61	HVA rating of the relevance of the information	85.3	97.2	>0.0[2]
62	Keywords describing content of the information, sometimes also time reference	99.2	99.1	100.0
63	Form of the information, e.g. report of IM (or responsible full-time HVA employee) or original documents (e.g. annual reports or meeting minutes)	99.2	99.2	100.0
64	**Inc.:** HVA department responsible for evaluating the incoming information **Out.:** The recipient of the outgoing information	99.6	99.2	73.1
65	Identifier(s) of information used to create the outgoing information	0.1	0.0	75.0
104	Identifier(s) of other information with contextual association to the information	0.0	0.0	0.0
106	**Inc.:** Identifier, code name and reliability of IM obtaining the incoming information **Out.:** Identifier, code name and reliability of IM(s) obtaining incoming information used for the outgoing information (before 1978)	70.3	77.0	24.1
107	Country of HVA field office involved in acquiring or transferring the information	10.9	12.4	0.0
108	**Inc.:** Continuous yearly counter of the incoming information (reset yearly)	87.4	99.6	0.0

Notes: This table lists all meta features on pieces of information recorded in the SIRA database which are available to us. The feature numbers are not consecutive because i) some features are not available for the part of SIRA data we analyze (subdatabase 12) and ii) we do not have acces to some features due to data protection. For each feature, we list the percentage of pieces of information (split by incoming and outgoing pieces) between 1969 and 1987 for which the value of the respective feature is non-missing. **1:** Some features are comprised by several subfeatures (e.g. feature 106 includes values on IMs' code names *and* identifiers). We classify a value of a feature as non-missing if any of the subfeatures for an piece of information is non-missing. Some features contain values which are encoded as missing values by the HVA but not recognized as such by our statistical software. For this table, we do not classify those as missing. For statistics on missing data in the main text, however, these values are considered as missing. **2:** The percentage of pieces of information with non-missing values is not zero but smaller than 0.05. **Source:** Konopatzky (2019).

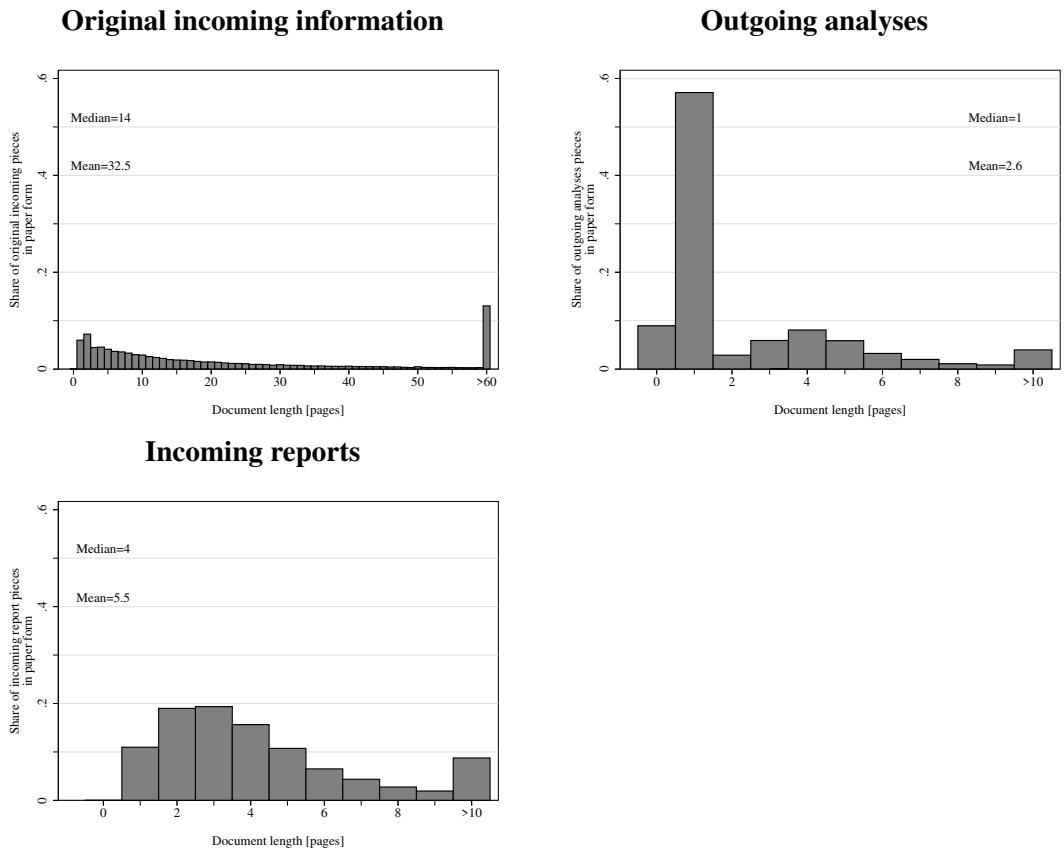
Table 3.A.2: Overview of available meta features, continued

Meta feature	Description	% of pieces of information with non-missing value[1]		
		Total	Inc.	Out.
110	Additional notes (standardized) regarding relevance of the information, e.g. in terms of novelty or credibility	>0.0	>0.0	0.0
111	Data medium on which the information is saved (includes paper)	100.0	100.0	100.0
112	Purpose for collecting or sending the information	66.1	74.2	8.0
113	**Out.:** Keywords (standardized) further describing type of the information (e.g. overview vs. evaluation)	12.2	0.0	100.0
115	**Out.:** Continuous yearly counter of the outgoing information (reset yearly)	10.6	>0.0	86.3
116	Keywords (standardized) describing broader thematic field of the information (only from mid September 1987)	5.4	5.4	8.5
117	Country to which the information is related	83.5	84.5	76.5
119	Institution or organization to which the information is related	54.8	55.6	49.3
121	Confidentiality of the information within the HVA	67.1	76.2	1.2
122	Additional notes regarding processing of the information (standardized)	1.4	1.0	4.5
123	Additional notes regarding processing of the information (free)	1.7	1.8	>0.0
124	Language of the information	8.8	10.0	>0.0
152	Additional notes on the information	18.6	18.6	18.5

Notes: This table lists all meta features on pieces of information recorded in the SIRA database which are available to us. The feature numbers are not consecutive because i) some features are not available for the part of SIRA data we analyze (subdatabase 12) and ii) we do not have acces to some features due to data protection. For each feature, we list the percentage of pieces of information (split by incoming and outgoing pieces) between 1969 and 1987 for which the value of the respective feature is non-missing. **1:** Some features are comprised by several subfeatures (e.g. feature 106 includes values on IMs' code names *and* identifiers). We classify a value of a feature as non-missing if any of the subfeatures for an piece of information is non-missing. Some features contain values which are encoded as missing values by the HVA but not recognized as such by our statistical software. For this table, we do not classify those as missing. For statistics on missing data in the main text, however, these values are considered as missing. **2:** The percentage of pieces of information with non-missing values is not zero but smaller than 0.05. **Source:** Konopatzky (2019).

3.A.1. Overview of pieces of information

Figure 3.A.1: Empirical distribution of document length for pieces of information in paper form



Notes: This figure shows the distribution of page numbers for different information types which were delivered to, or sent out by, the HVA in paper form. Page number values are censored at 60, 10 and 10 pages, respectively.

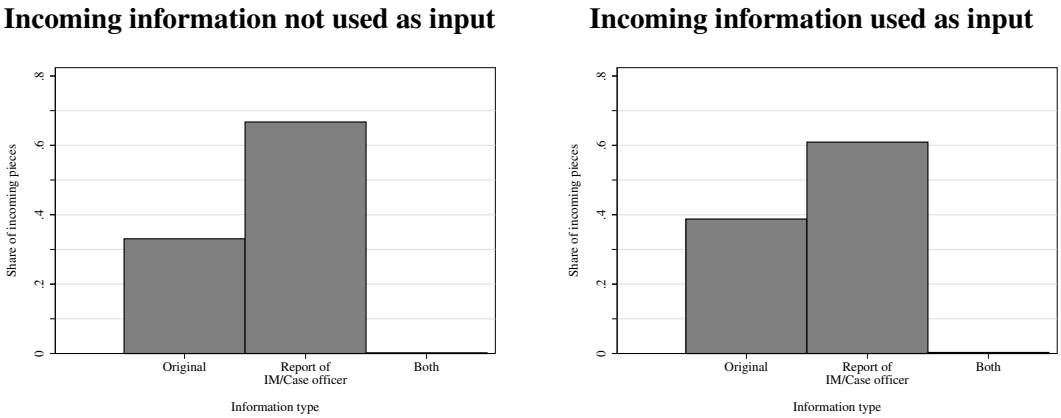
Table 3.A.3: Most frequent country references

Country		% of pieces labelled		
		Incoming	Outgoing	(rank)
1	Central Europe	56.5	66.5	(1)
2	West Germany	46.6	57.3	(2)
3	America	31.2	31.0	(3)
4	USA	29.9	30.7	(4)
5	Far East	19.4	17.1	(8)
6	USSR	15.7	19.8	(6)
7	Eastern Europe	15.1	19.4	(7)
8	Middle East	15.0	11.0	(10)
9	Western Europe	14.7	16.7	(9)
10	GDR	14.0	22.6	(5)
11	Northern Africa	8.5	6.4	(16)
12	West Berlin	8.4	10.5	(11)
13	China	8.2	8.1	(13)
14	Southern Europe	7.4	7.7	(14)
15	France	6.7	8.8	(12)
16	Great Britain	5.3	7.0	(15)
17	Israel	4.6	3.5	(20)
18	Egypt	4.2	3.1	(22)
19	Iran	4.2	2.7	(24)
20	Syria	4.2	3.4	(21)
21	Poland	3.8	4.5	(18)
23	Europe	3.6	6.3	(17)
24	Japan	3.6	4.2	(19)

Notes: This table shows the percentage of pieces of information labelled by the 20 most frequently used country descriptors for incoming pieces. The percentage and rank (in parentheses) of the same country descriptors among outgoing pieces of information is also given. The last 3 country descriptors are shown because they are among the 20 most frequently used descriptors for outgoing pieces of information.

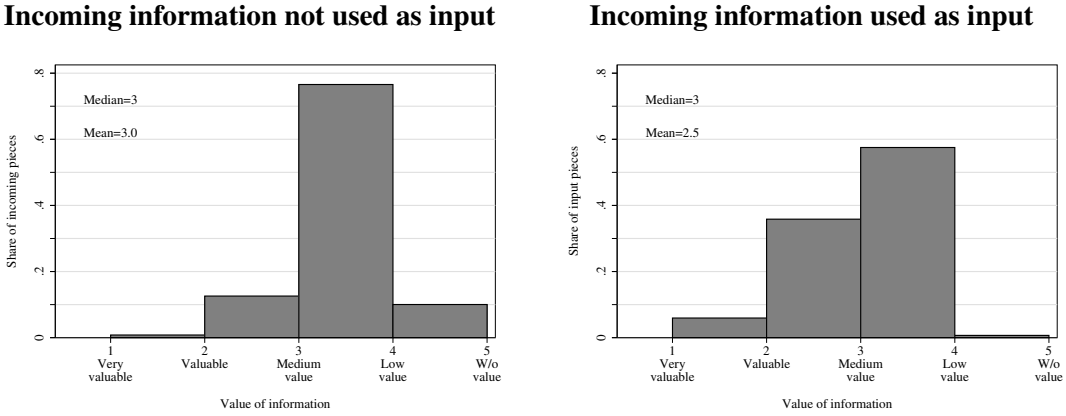
3.A.2. Linking pieces of information

Figure 3.A.2: Comparison of empirical distribution of information form



Notes: This figure compares the distribution of information types over pieces of information between incoming pieces used and not used as input for outgoing pieces. Only pieces of information dated between 1983 and 1987 are considered.

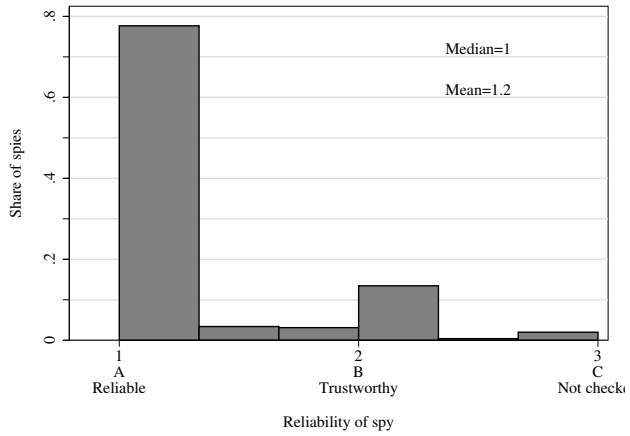
Figure 3.A.3: Comparison of empirical distribution of relevance of pieces of information



Notes: This figure compares the distribution of information relevance between incoming pieces used and not used as input for outgoing pieces. Since the values vary per piece of information, the figure shows average values per piece of information. Only pieces of information dated between 1983 and 1987 are considered.

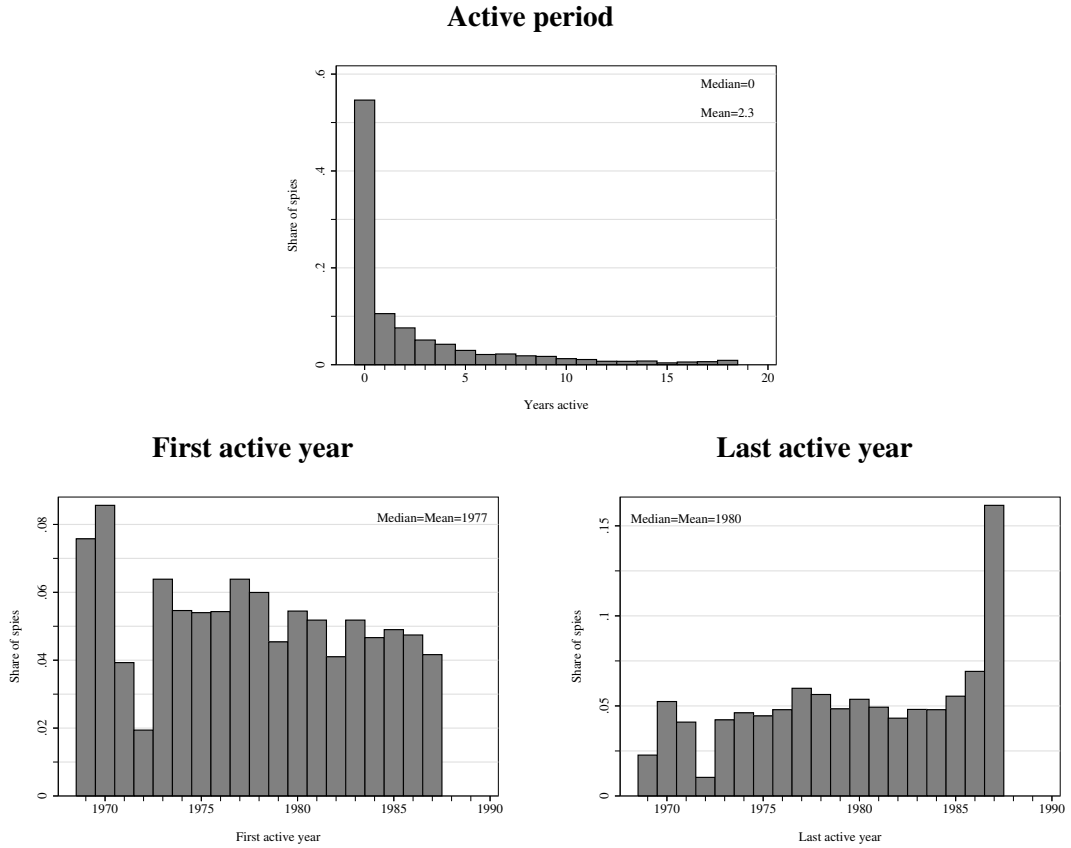
3.A.3. Informants

Figure 3.A.4: Empirical distribution of spies' reliability



Notes: This figure shows the distribution of reliability scores over spies. The HVA rated reliability on an ordinal scale with values A (*Reliable*), B (*Trustworthy*), C (*Not checked*), D (*Questionable*) and E (*Double agent*). Since reliability scores vary at the piece of information level, the plot is based on the average (across all pieces of information gathered by a spy) of the encoded (A=1, B=2, C=3) recorded assessments.

Figure 3.A.5: Empirical distributions of spies' active period and first as well as last active year



Notes: This figure shows the distributions of the length of active periods, of first and last active years over spies. The active period of a spy is the difference between the dates of her first and last information provided.

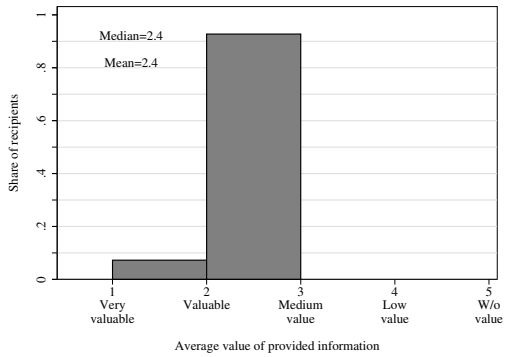
Table 3.A.4: Most frequent institutional references of top 20 spies and recipients

Institution/Organization	Translation/Explanation
BDI	*Bundesverband der Deutschen Industrie* (Federation of German Industries)
EG	*Europäische Gemeinschaft* (European Community)
U-VERBAND	*Unternehmerverband* (Employer Association)
CDU	Christian Democratic Union of German (Christian conservative party)
SPD	Social Democratic Party of Germany
FDP	Free Democratic Party (liberal party)
PLO	Palestine Liberation Organization
KSZE	*Konferenz über Sicherheit und Zusammenarbeit in Europa* (Conference on Security and Co-operation in Europe)
CNAD	Conference of National Armaments Directors (Nato conference)
BUNDESWEHR	West German armed forces
WV	N/A
DRG	N/A
NAFAG	NATO Air Force Armaments Group
AA	N/A
SENAT	State Cabinet of West Berlin
BUNDESTAG	West German federal parliament
DPC	N/A
WELTBANK	World Bank
FES	*Friedrich-Ebert-Stiftung* (Friedrich-Ebert-Foundation)[1]
ABGEORDNETENHAUS	State parliament of West Berlin
BMFT	*Bundesministerium für Forschung und Technologie* (West German ministry for Research and Technology)
IWF	*Internationaler Währungsfonds* (International Monetary Fund)
COCOM	Committee on Multilateral Export Controls
RGW	*Rat für gegenseitige Wirtschaftshilfe* (Socialist pendant of OECD)

Notes: This table shows translations and explanations of the institutional references listed in Tables 3.3.2 and 3.3.3. "N/A" in column 2 indicates that we do not know to which organization or institution the descriptor in column 1 refers. **1:** Traditionally, each West German party maintains a foundation which, amongst other purposes, engages in political education and maintaining its party's archives. For legal reasons these foundations are separate legal entities. The Friedrich-Ebert-Foundation is the foundation of the SPD.

3.A.4. Recipients

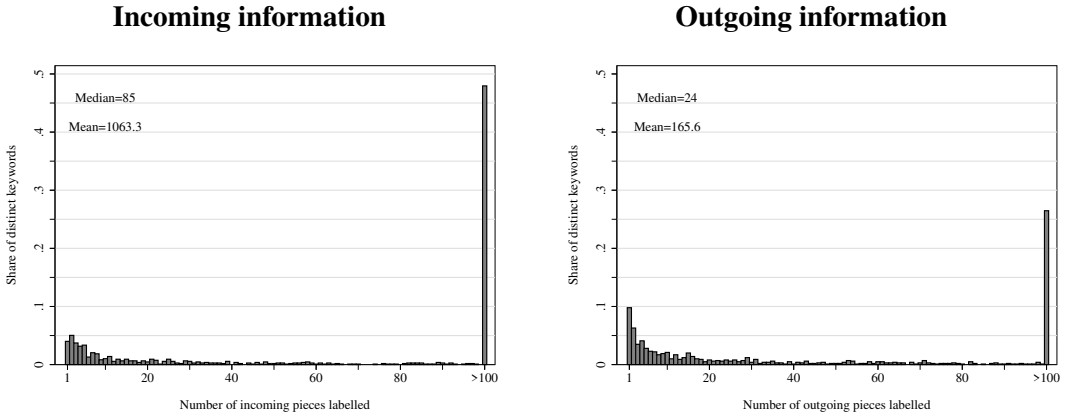
Figure 3.A.6: Empirical distribution of recipient's relevance of received information



Notes: This figure shows the distribution of the average relevance scores over recipients. Relevance is measured by the average relevance score of the input pieces used per outgoing piece. The HVA rated relevance on an ordinal scale with values 1 (*Very valuable*), 2 (*Valuable*), 3 (*Medium value*), 4 (*Low value*) and 5 (*Without value*). Based on outgoing information only.

3.A.5. Keywords

Figure 3.A.7: Empirical distribution of the frequency of keywords in the data



Notes: This figure shows the distribution of keyword frequencies for incoming and outgoing pieces of information. For better readability, all instances of keywords occurring more often than 100 times are shown in the rightmost bin. The average density of these keywords is 0.001 for both plots.

3.B. Appendix to Section 3.5

Table 3.B.1: Manual changes to keywords

Chosen version	Version 1	Version 2	Version 3	Version 4	Version 5
AKP-STAAT	AKP-STAATEN				
ANLAGE	ANLAGEN				
BERICHT	BERICHTE				
BEZIEHUNG	BEZIEHUNGEN				
EINSCHAETZUNG	EINSCHAETZUNGEN				
ELEKTRONISCHE DATENVERARBEITUNGS-ANLAGEN UND SYSTEME	ELEKTRONISCHE DATENVERARBEITUNGS-ANLAGE				
ENTWICKLUNGSTENDENZ	ENTWICKLUNGSTENDENZEN				
FLUECHTLINGSPROBLEM	FLUECHTLINGSPROBLEME				
GEPANZERTES FAHRZEUG	GEPANZERTE FAHRZEUGE				
GERAET	GERAETE				
GESPRAECH	GESPRAECHE				
GEWERKSCHAFT	GEWERKSCHAFTEN				
INVESTITION	INVESTITIONEN				
KATH.KIRCHE	KATHOLISCHE KIRCHE				
KERNBRENNSTOFF	KERNBRENNSTOFFELEMENT				
KERNFORSCHUNG	KERNFORSCHUNGSZENTRUM	KERNTECHNIK			
KERNWAFFE	KERNWAFFEN				
KONTAKT	KONTAKTE				
KREDIT	KREDITE	KREDITGEWAEHRUNG			
KRIEGSDIENSTVERWEIGERER	KRIEGSDIENSTVERWEIGERUNG				
LAGE	LAGEEINSCHAETZUNG				
LANDSMANNSCHAFT	LANDSMANNSCHAFTEN				
LOHNPOLITIK	LOHN- UND TARIFPOLITIK				
LUFTLANDETRUPPE	LUFTLANDETRUPPEN				
LUFTSTREITKRAEFTE	LUFTWAFFE				
MASZNAHME	MASZNAHMEN				

Notes: This table shows manual changes performed to keywords. Each column represents a different version of - as we assume - the same keyword. We replace versions in columns 2 to 6 with the version shown in column 1.

Table 3.B.2: Manual changes to keywords, continued

Chosen version	Version 1	Version 2	Version 3	Version 4	Version 5
MATERIELLE MOBILMACHUNG	MATERIELLE MOBILISIERUNG				
MENSCHENRECHTE	MENSCHENRECHT				
MILITAERISCHE FORDERUNGEN	MILITAERISCHE FORDERUNG				
MILITAERTRANSPORT	MILITAERTRANSPORTWESEN				
MOBILMACHUNG	MOBILISIERUNG				
NAHOSTPOLITIK	NAHOST-POLITIK				
OST-WEST-BEZIEHUNG	OST-WEST-BEZIEHUNGEN				
PARTEITAG	PARTEITG				
PERSONELLE MOBILMACHUNG	PERSONELLE MOBILISIERUNG				
ROHSTOFF	ROHSTOFFE				
SCHWIERIGKEIT	SCHWIERIGKEITEN				
STABSUEBUNG	STABSUEBUNGEN				
STRATEGISCHES MANOEVER	STRATEGISCHES MANOEVER/ UEBUNG	STRATEGISCHE UEBUNG	STRATEGISCHE MANOEVER/ UEBUNG	MANOEVER UND UEBUNGEN	MANOEVER
STREITKRAEFTEPLANUNG	STREITKRAEFTEPLAN				
TERRORISMUS	TERROR				
UEBUNG	UEBUNGEN				
UMSTRUKTURIERUNG	UMSTRUKTUIERUNG				
UNIVERSITAET	HOCHSCHULEN UND UNIVERSITAETEN	HOCHSCHULEN			
UNO-SPEZIALORGANISATIONEN	UNO-SPEZIALORGANISATION				
UNTEROFFIZIER	UNTEROFFIZIERE				
VERTRAUENSBILDENDE MASZNAHMEN	VERTRAUENSBILDENDE MASZNAHME	VERAUENSBILDENDE MASZNAHMEN			
WAFFE	WAFFEN				
WAHLERGEBNIS	WAHLBUENDNISSE				
WEHRERSATZ	WEHRERSATZWESEN				
WEHRGESETZGEBUNG	WEHRGESETZWESEN				
WIEDERAUFARBEITUNG VON KERNBRENNSTOFF	WIEDERAUFARBEITUNG VON KERNBRENNSTOFFEN	WIEDERAUFBEREITUNG VON KERNBRENNSTOFF			

Notes: This table shows manual changes performed to keywords. Each column represents a different version of - as we assume - the same keyword. We replace versions in columns 2 to 6 with the version shown in column 1.

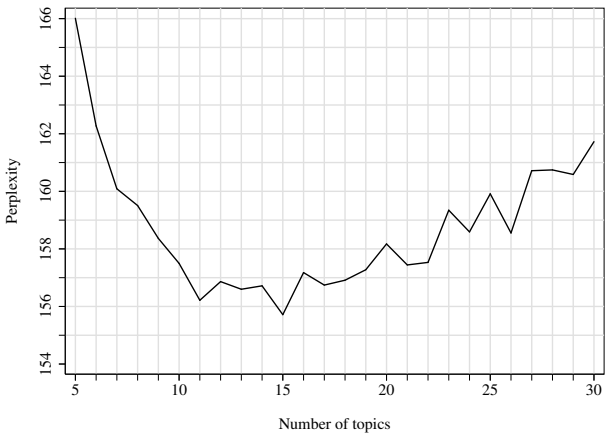
3.B.1. Model selection

To chose the appropriate LDA model for our semantic analysis, i.e. determine the number of topics T and values of hyper parameters α and η, we cross-validate results from different models. Using each model's perplexity value as goodness-of-fit measure, we select the one with the lowest score. Perplexity is a negative function of the likelihood to observe a particular corpus of documents and words given a previously estimated LDA model. It has been widely applied for model selection purposes in semantic analysis (Blei, Ng, and Jordan, 2003; Griffiths and Steyvers, 2004).

In many applications of LDA, for model selection, the hyper parameters α and η are fixed and then T is chosen such that it minimizes perplexity. As a heuristic, $\alpha = \frac{50}{T}$ and $\beta = 0.1$ are often chosen as values for the hyper parameters (Griffiths and Steyvers, 2004; Hansen, McMahon, and Prat, 2018; Meyer et al., 2019; Lüdering and Tillmann, 2020). However, Meyer et al. (2019) note that for corpora with relatively short documents, in particular Twitter posts, this heuristic is inappropriate. It is likely that, for short documents, the number of topics occurring per document is considerably smaller than for long documents. Therefore, Meyer et al. (2019) suggest smaller values of α, ranging between $\alpha = \frac{1}{2T}$ and $\alpha = \frac{5}{T}$. To consider this intuition, we estimate LDA models with α taking values $\alpha = \frac{1}{4T}$, $\alpha = \frac{1}{2T}$, $\alpha = \frac{1}{T}$, $\alpha = \frac{2}{T}$ and $\alpha = \frac{5}{T}$. For η, the authors suggest values ranging from $\eta = 0.001$ to $\eta = 0.01$, implying that, in short documents, topics are comprised by fewer important words than in long documents. While Meyer et al. (2019) do not explain this parameter choice, it does make sense for our data: keywords constitute a type of label for pieces of information and were probably chosen such that they describe an information's content as accurately and shortly as possible. A single keyword describes the same content as many single "normal" words, albeit less exactly. Therefore, it is likely that single *keywords* are more important for a certain topic, i.e. they have a high probability mass in the distribution over words which constitutes a topic, than single normal words would be. Because of these considerations, we also vary the values of η in the same way as we do for α. For the range of topics we consider (recall that the particular values of α and η depend on T), the respective values of η range between $\eta \approx 0.01$ and $\eta = 1$.

Finally, for each of the 25 possible combinations of parameter values for α and η, we estimate LDA models using $T = 5$ to $T = 30$ topics. We follow Hansen, McMahon, and Prat (2018) and fit the LDA models for each of these 650 α-η-T combinations on a randomly drawn, two-third subsample of our pre-processed data. Then we calculate the respective perplexity value on the remaining third of the data. We iterate this procedure 10 times. Figure 3.B.1 plots the results of this analysis. For each topic and iteration, we collect the α-η combination which minimizes perplexity. The plot in Figure 3.B.1

Figure 3.B.1: Perplexity as a function of number of topics



Notes: This figure shows, for each number of topics, the average, across 10 iterations, of the minimum perplexity, obtained across all possible combinations of hyper parameters α and η.

associates each number of topics with its across-iteration average, minimum perplexity value.

The graph indicates that, with respect to perplexity, the optimal number of topics is 15. We chose the remaining hyper parameters by averaging the values of α and η associated with the minimum perplexity value per cross-validation iteration across iterations. Using 15 topics, however, we find that there is some semantic overlap between topics. Next to goodness-of-fit, interpretability of results should be another important factor for model selection (Chang et al., 2009, Hansen, McMahon, and Prat, 2018). Therefore, we also estimate models with 11, 12, 13 and 14 topics, respectively. That way we remove some of the overlapping topics while still ensuring very low perplexity values. Additionally, for expositional convenience, we also fit a model with 10 topics. Tables 3.B.3 to 3.B.8 show the respective outcomes by listing, for each topic, the 15 most important words.

Results For the remaining paragraph, we distinguish between the terms *Topic* and *Subject*. We will use the term topic as defined in the LDA context and use the term subject to refer to sets of keywords which we identify to be semantically close. We use this distinction because we find that, while reducing the number of topics to be identified by LDA, keywords that would make up stand-alone topics in LDA models allowing for relatively many total topics will appear within one, merged, topic when allowing only relatively few total topics. For example, estimating the LDA model with 15 topics in total yields two topics which we label *Military strategy* and *Military technology*, respectively. For models with fewer topics, e.g. 11, we find that the keywords defining

topics *Military strategy* and *Military technology* in the LDA model with 15 topics, appear together in one topic. Therefore, to avoid confusion, we refer to the semantic unit of keywords that we identify and label as *Military strategy* or *Military technology* as subjects. The keywords identifying these subjects can manifest as two topics within an LDA model or can appear together with other keywords characteristic for other subjects.

There are three subjects which emerge as one topic in all specifications. Irrespective of our hyper parameter choices, the keywords associated with the highest probability masses within the distributions that constitute a topic barely vary. First, we label the topic identified by keywords *State apparatus*, *Government*, *Relation*, *Assessment* and *Stance* as *Relation to West German government*. The preceding keywords appear among the top words for this topic across all specifications. Other important keywords are *East-West relation*, *Domestic policy* and *Journey*. The descriptors are mostly self-explanatory: pieces labelled by a combination of these keywords likely contain information on the stance of the West German government regarding different policy areas. The descriptor *Journey* probably refers to diplomatic visits of East and West German government officials in the respective other country.

The second topic for which the most important words are stable across specifications describes the *Political landscape in West Germany*. Keywords *Party*, *Election*, *Differences*, *State of interior affairs* and *Coalition* appear within the list of top 15 words in all models. Furthermore, keywords *Election campaign*, *SP* (probably Socialist Party), *Parliament* and *Party congress* are often associated with very high probability mass. pieces of information labelled with a combination of these descriptors likely contained details on West German parties, their relationships with each other and on parliaments. The Stasi espionage focused on the West German federal parliament, uncovered by BStU (2013a), is probably also attributable to this topic.

Third, we find that keywords *Security policy*, *East-West relation*, *Stance*, *Disarmament*, *Armament* and *Negotiation* form a stable topic which we decide to label *Geo & security politics*. Further frequently occurring descriptors are *MVM*, *Conference*, *KSZE* and *Detente*. *KSZE* is the German abbreviation for the Conference on Security and Co-operation in Europe (CSCE), a series of summits between the NATO and the states of the Warsaw Pact integral to the detente process during the Cold War. Keyword *MVM*, the abbreviation for *Militärverbindungsmission*, Military liaison missions in English, describes regular military missions which were intended to allow representatives of Western and Soviet military intelligence to monitor each other's occupied parts of Germany.

Two further subjects appear in one topic each for almost all models. The first we label as *Trade with West Germany*. Its most important keywords are *Trade*, *Export*,

Import, *(East-West) relation* and *Commodity*. The second subject we label *Civil R&D (energy)*. It is characterized by keywords *Energy*, *Nuclear energy*, *Industry*, *Energy policy*, *Energy economy*, *Nuclear research*, *Nuclear industry* and *Crude oil*. When estimating an LDA model with 10 topics in total, both subjects are merged into one topic which we then call *Trade & Energy*. For the LDA model with 11 total topics we find that trade related keywords appear in one topic with descriptors of a subject we refer to as *Economic relation with West Germany*. Accordingly we label the respective topic as *Trade & Economic relation with West Germany*.

The subject we decide to label *Economic relationship with West Germany* is characterized by keywords *Economy*, *Economic policy*, *Collaboration*, *Relation*, *Financial affairs*, *Development policy*, *Industry*, *Conference*, *Investment*, *Development aid* and *Entrepreneur's association*. The subject manifests in at least one topic for all specifications except for the model with 11 total topics. Here, as mentioned previously, some of its keywords can be found within the topic *Trade & Economic relation with West Germany*. Allowing for 10, 13 or 14 total topics, the subject's keywords additionally appear in one further topic together with keywords of the subject *Trade with West Germany* (10 total topics) or of a subject we label *Civil R&D* (13 and 14 total topics). We refer to these topics as *Trade & economic relationship with West Germany* and *Economic & scientific relation with West Germany*, respectively. In the specification with 14 total topics, descriptors of the subject *Economic relationship with West Germany* further appear together in one topic with keywords of a subject related to military espionage. We call the respective topic *Economics & military*. Finally, allowing for 15 total topics, the subject manifests in two distinct topics without being mixed with other subjects. Here, keywords *Firm*, *Financing*, *Payment transactions*, *Credit* and *Budget* are further characteristic identifiers.

Keywords of the previously mentioned subject *Civil R&D* appear for models with 11 to 15 total topics. These are *Science*, *Research*, *Development* and *Firm*. For models with 11 and 12 total topics, the subject appears together with descriptors of military espionage such that we label the respective topics *Civil R&D & military technology*. As described in the preceding paragraph, for models with 13 and 14 total topics, the subject's keywords occur together in one topic with descriptors of the subject *Economic relationship with West Germany*. For 15 total topics, the subject *Civil R&D* manifests as singular topic.

Another two very important subjects are concerned with military espionage, labelled either *Military strategy* or *Military technology*. While the former subject is identified by keywords *Military affairs*, *Military policy*, *Land forces*, *Military strategy*, *Air force*, *Exercise*, *Command*, *Defense* and *Naval forces*, the latter features descriptors like *Military*

technology, *Research*, *Development*, *Armament*, *Project* and more concrete keywords like *Missile*, *Airplane* or *Weapon*. Across different models, the subject *Military strategy* manifests as one single topic (10, 12, 13 and 15 total topics) or mixed with keywords of a subject we label *International conflicts* (10, 11, 13 and 14 total topics). Subject *Military technology* emerges as stand-alone topic (10, 13 and 15 total topics) or together with keywords of the subject *Civil R&D* (11 and 12 total topics). For the model with 13 total topics we additionally label one topic as *Military technology (intelligence)* because the respective keywords appear alongside with descriptors *Intelligence* and *MVM*. Sometimes both military subjects also appear together in one topic (11 and 14 total topics). Finally, as already mentioned, for the specification allowing 14 total topics we find one topic that is identified by keywords of both military related subjects and the *Economic relation with West Germany* subject.

Closely related to the subject of military espionage is a set of keywords which we associate with the label *International conflicts*. These keywords are *Relation*, *Differences*, *Conflict settlement*, *Stance*, *Domestic policy*, *Middle-East policy*, *Conflict*, *War*, *Terrorism*, *Liberation movement*. The subject's descriptors appear together with those of the subject *Military strategy* for models with a total of 10, 11, 13 and 14 topics. In the model with a total of 12 topics some of these relevant keywords are found within the *Geo & security politics* topic whereas they appear without another subject's keywords in the model with a total of 15 topics. In that case we label the respective topic *International Relations & conflicts*.

Another set of keywords define a subject that we label as *Enemy activity*. The respective topics all feature the keyword *Enemy activity*, *Feindtätigkeit* in German, an established term within the MfS (Müller-Enbergs, n.d.). Most topics we associate with this subject also exhibit the descriptor *PID* as important keyword. PID is an abbreviation for *Politisch-ideologische-diversion* (Political-ideological-deviation), also a very common part of the Stasi terminology which was subsequently adopted by many communist security services (Engelmann, n.d.). The term was used to label the ideological influence of the so-called West and smaller entities (e.g. institutions or individuals whose ideology deviated from the SED's official doctrine) on communist societies. Most often (for models with a total of 11, 12 and 15 topics) we observe that the subject is distributed over at least two topics: when the above mentioned keywords are accompanied by descriptors *Population*, *Church* or *Youth*, we label the associated topic as *Enemy activity (civil society)*. Those topics additionally featuring descriptors *Enemy organisation*, *Secret service* or *State apparatus*, we label as *Enemy activity (foreign nations & intelligence)*. For models with a total of 10 and 13 topics all of the above keywords appear together in one topic along with descriptors indicating the involvement of the me-

dia. These are *Media*, *Press* and *Information technology*. We therefore opt for a more general label and chose *Enemy activity*. In the LDA model with 11 total topics we find that next to topics *Enemy activity (foreign nations & intelligence)* and *Enemy activity (civil society)* there is a topic which we can label *Enemy activity (media)*. For the model with 14 total topics we can only identify the topic *Enemy activity (foreign nations & intelligence)*, however we do find the media and civil society related keywords appearing together in one topic which we then label *Civil society & media*. In the same model we decide to label one topic as *Civil society in West Germany* as it contains keywords like *Organisation*, *Extremism*, *The Left*, *The Right*, *Union* and *Maoism*. Finally, for the model with a total of 15 topics we only identify the *Enemy activity (foreign nations & intelligence)* and *Enemy activity (civil society)* topics. However, we find that the media related keywords appear together in a topic with descriptors previously associated with the topic *Political landscape in West Germany*. We therefore label the respective topic as *Political & media landscape in West Germany*.

Table 3.B.3: Top words per topic, 10 topics

Relation to West German government	**Political landscape in West Germany**	**Geo & security politics**	**Trade & Energy**	**Trade & economic relationship with West Germany**
State apparatus	Party	Security politics	Trade	Economy
Government	Election	East-West relation	Payment transactions	Trade
Relation	Domestic policy	Disarmament	Energy	Collaboration
Assessment	State of interior affairs	Armament	Industry	KIL
Domestic policy	Differences	Stance	Economic policy	SL
Differences	Assessment	MVM	Nuclear energy	Relation
Journey	Parliament	Nuclear weapon	Loan	East-West relation
Stance	KP	KSZE	Energy policy	Industry
Party	Election campaign	Negotiation	Export	Economic policy
Treaty	Coalition	Conference	Import	WIA
Transportation	Organisation	Detente	Assessment	EL
West Berlin status	SP	Party	Energy economy	Commodity
East-West relation	Party congress	SL	East-West relation	Development policy
Ostpolitik	Differentiation	Assessment	Economy	Firm
Diplomatic representation	State of affairs	Check	Collaboration	Assessment
Economic relationship with West Germany	**International conflicts & military strategy**	**Military technology**	**Military strategy**	**Enemy activity**
Economy	Domestic policy	Military technology	Military affairs	Enemy activity
Economic policy	Differences	Armament	Armed forces	PID
Assessment	Relation	Military affairs	Military policy	Contact
Collaboration	Conflict settlement	Armed forces	Assessment	Media
Trade	Stance	Development	Land forces	Information technology
Relation	Military affairs	Electronics	Military strategy	Organisation
Financial concerns	State of interior affairs	Research	Structure	Relation
Development policy	Middle-East policy	Collaboration	Planning	Church
Industry	Armed forces	Missile	Air force	Enemy organisation
Conference	Liberation movement	Project	Exercise	Measure
Development	Opposition	ELoKa	Staffing	Ostforschung
Domestic policy	Assessment	Industry	Military technology	Press
Entrepreneur association	Conflict	Aircraft	Command	Idelogical diversion
Investment	War	Weapon	Defense	Intelligence agency
Development aid	Terrorism	Air force	Naval forces	Conference

Notes: This table shows the English translation of the 15 most important keywords for each of 10 topics identified by an LDA model with $\alpha = 0.043333333$ and $\eta = 0.038333333$.

Table 3.B.4: Top words per topic, 11 topics

Relation to West German government	**Political landscape in West Germany**	**Geo & security politics**	**Trade & Economic relation with West Germany**	**Civil R&D (energy)**
State apparatus	Party	Security politics	Trade	Commodity
Government	Election	East-West relation	Economy	Energy
Assessment	Domestic policy	Disarmament	Collaboration	Nuclear energy
Relation	State of interior affairs	Armament	Economic policy	Economic policy
Domestic policy	Differences	Stance	Relation	Industry
Differences	Parliament	Negotiation	Development policy	Energy policy
Stance	Assessment	Detente	Assessment	Crude oil
Journey	KP	KSZE	Payment transactions	Energy economy
Party	SP	Conference	Financial concerns	Economy
Coalition	Election campaign	SL	Industry	Assessment
East-West relation	Coalition	Assessment	East-West relation	Research
State of interior affairs	Party congress	MVM	Export	Collaboration
Election	Command	Party	Loan	Environmental protection
Conception	Organisation	Nuclear weapon	Agriculture	Nuclear research
Economic policy	Union	Check	Investment	Nuclear industry
Civil R&D & military technology	**Military strategy & technology**	**Military strategy & international conflicts**	**Enemy activity (foreign nations) (& intelligence)**	**Enemy activity (civil society**
Research	Armed forces	Relation	Enemy activity	Enemy activity
Armament	Military affairs	Differences	Security	Domestic policy
Development	Military technology	Military affairs	Enemy organisation	Difficulty
Industry	Military policy	Stance	Intelligence agency	Relation
Electronics	Land forces	Conflict settlement	Counter revolution	State of affairs
Collaboration	Air force	Middle-East policy	Military affairs	Assessment
Firm	Armament	Domestic policy	Law	SL
Military technology	Assessment	Armed forces	Terrorism	Population
Project	Planning	Military policy	State apparatus	Economy
IT	Reconnaissance	State of interior affairs	Domestic policy	PID
Financing	Naval forces	Liberation movement	Measure	Ostforschung
Economy	Collaboration	Conflict	State of interior affairs	Church
Science	Military strategy	Assessment	Main problem	East-West relation
Electrotechnology	ELoKa	War	Extremism	Situation
Report	Missile	Arms supply	ZV	Youth
Enemy activity (Media)				
Enemy activity				
PID				
Relation				
Contact				
Media				
Information technology				
West Berlin status				
Transportation				
Organisation				
Treaty				
Ostpolitik				
Stance				
Socialist countries				
Press				
Assessment				

Notes: This table shows the English translation of the 15 most important keywords for each of 11 topics identified by an LDA model with $\alpha = 0.03125$ and $\eta = 0.0328125$.

Table 3.B.5: Top words per topic, 12 topics

Relation to West German government	**Relation to West Germany**	**Political landscape in West Germany**	**Civil society**	**Geo & security politics**
State apparatus	Party	Election	Organisation	Security politics
Government	Domestic policy	Party	Extremism	Disarmament
Relation	State of interior affairs	Parliament	Left	Stance
East-West relation	Differences	Election campaign	Structure	East-West relation
Assessment	State apparatus	Coalition	Population	Armament
Journey	Government	Differences	Law	Negotiation
Stance	Command	Assessment	Measure	Relation
Treaty	KP	State of interior affairs	Right	Conflict settlement
West Berlin status	Assessment	SP	Security	Middle-East policy
Differences	Economic policy	Differentiation	Maoism	Differences
Ostpolitik	Stance	Parliamentary faction	Terrorism	Military policy
Discussion	SP	Strategy	Youth	Conference
Diplomatic representation	Union	Tactic	Domestic policy	Assessment
Detente	Opposition	State parliament	Conference	MVM
Socialist countries	Social policy	Domestic policy	Assignment	Conflict
Economic relation with West Germany	**Trade with West Germany**	**Civil R&D (energy)**	**Military technology & civil R&D**	**Military strategy**
Development policy	Trade	Industry	Research	Military affairs
Relation	Economy	Energy	Armament	Armed forces
Collaboration	Economic policy	Nuclear energy	Development	Military policy
Economy	Payment transactions	Economic policy	Electronics	Military technology
Financial concerns	Relation	Energy policy	Industry	Land forces
Trade	Export	Assessment	Collaboration	Air force
Conference	East-West relation	Economy	Military technology	Assessment
Economic policy	Loan	Transportation	Firm	Armament
Assessment	Collaboration	Energy economy	Project	Planning
Development aid	Assessment	State of affairs	Science	Reconnaissance
Project	Commodity	Trade	IT	Collaboration
Financial policy	KIL	Research	Economy	Military strategy
Treaty	SL	Firm	Financing	Naval forces
Agriculture	Crude oil	Collaboration	Electrotechnology	ELoKa
Development countries	Import	Commodity	Committee	Exercise
Enemy activity (foreign nations & intelligence)	**Enemy activity (civil society)**			
Enemy activity	Enemy activity			
East-West relation	PID			
SL	Relation			
Enemy organisation	Domestic policy			
Ostforschung	Contact			
Conference	Media			
menschenrechte	Information technology			
KSZE	Difficulty			
Intelligence agency	Church			
Assessment	Party			
PID	Organisation			
SSG	Counter revolution			
German question	Press			
Soc/ analysis	Differences			
Influence	Idelogical diversion			

Notes: This table shows the English translation of the 15 most important keywords for each of 12 topics identified by an LDA model with $\alpha = 0.03529411764705882$ and $\eta = 0.057352941176470586$.

Table 3.B.6: Top words per topic, 13 topics

Relation to West German government	**Relation to West Germany**	**Political landscape in West Germany**	**Geo & security politics**	**Trade with West Germany**
State apparatus	Domestic policy	Party	Security politics	Trade
Government	Party	Election	East-West relation	Economy
Relation	Relation	State of interior affairs	Disarmament	East-West relation
Assessment	Differences	Parliament	Stance	SL
Journey	State apparatus	Domestic policy	Armament	KIL
Treaty	KP	Differences	KSZE	Export
Stance	Government	Coalition	Detente	Enemy activity
West Berlin status	State of affairs	Assessment	Negotiation	Economic policy
Differences	Difficulty	Election campaign	Conference	WIA
Ostpolitik	Assessment	Differentiation	Party	Assessment
Conference	Economy	Party congress	Assessment	Collaboration
Socialist countries	Opposition	SP	SL	Relation
East-West relation	Regime	Stance	Measure	Import
Collaboration	Stance	Economic policy	friedensbewegung	Commodity
Discussion	SP	Union	Check	Embargo
Economic & scientific relation with West Germany	**Economic relation with West Germany**	**Civil R&D (energy)**	**Military strategy & international conflicts**	**Military technology**
Economy	Trade	Energy	Conflict settlement	Armament
Industry	Development policy	Nuclear energy	Middle-East policy	Military technology
Entrepreneur association	Payment transactions	Industry	Relation	Development
Assessment	Economy	Energy policy	Differences	Collaboration
Collaboration	Relation	Economic policy	Stance	Military affairs
Science	Collaboration	Commodity	State of interior affairs	Research
Research	Loan	Energy economy	Military affairs	Project
Economic policy	Economic policy	Crude oil	Military policy	Armed forces
Development	Financing	Research	Conflict	Aircraft
Firm	Financial concerns	Assessment	Armed forces	Industry
Financial concerns	Law	Nuclear research	War	Weapon
Financing	Transportation	Nuclear industry	Assessment	Firm
Monopoly	Development aid	Power plant	Liberation movement	Military policy
Committee	Assessment	Collaboration	Influence	Financing
Investment	Treaty	Environmental protection	Enemy activity	Aviation industry
Military technology (intelligence)	**Military strategy**	**Enemy activity**		
Military technology	Military affairs	Enemy activity		
Military affairs	Armed forces	PID		
Electronics	Military policy	Organisation		
MVM	Land forces	Contact		
Missile	Assessment	Media		
Armed forces	Structure	Information technology		
Nuclear weapon	Air force	Extremism		
ELoKa	Planning	Relation		
Reconnaissance	Exercise	Enemy organisation		
Armament	Staffing	Assessment		
Electrotechnology	Military strategy	Press		
Signal communications	Command	Population		
Assessment	Education	Party		
Deployment	Naval forces	Collaboration		
Electronic warfare	Plan	Measure		

Notes: This table shows the English translation of the 15 most important keywords for each of 13 topics identified by an LDA model with $\alpha = 0.03055555$ and $\eta = 0.03055555555$.

Table 3.B.7: Top words per topic, 14 topics

Relation to West German government	**Relation to West Germany**	**Political landscape in West Germany**	**Civil society in West Germany**	**Civil society & media**
State apparatus	East-West relation	Election	Organisation	Media
Government	Relation	Party	Domestic policy	Information technology
Domestic policy	KSZE	Parliament	Extremism	Church
Party	Transportation	State of interior affairs	Left	Press
Relation	West Berlin status	Election campaign	Social policy	FF
Differences	Stance	Coalition	Union	Public relations
KP	State apparatus	Assessment	Party	Program
Assessment	Treaty	Differences	Law	Measure
State of interior affairs	Government	Differentiation	Right	Culture
Stance	Journey	Population	Differences	Cath/ church
SP	Assessment	Party congress	Assessment	Assignment
Command	Conference	State parliament	Economic policy	Organisation
Economy	Detente	Tactic	Maoism	Conference
Difficulty	German question	Strategy	Measure	Collaboration
Opposition	Ostpolitik	Youth	Terrorism	Financing
Geo & Security politics	**Economic relation with West Germany**	**Economic & scientific relation with West Germany**	**Trade with West Germany**	**Civil R&D (energy)**
Security politics	Development policy	Economy	Trade	Economic policy
Disarmament	Relation	Trade	Economy	Commodity
East-West relation	Collaboration	Industry	Relation	Energy
Armament	Financial concerns	East-West relation	Export	Nuclear energy
Negotiation	Conception	Firm	Economic policy	Energy policy
Stance	Project	KIL	Collaboration	Industry
MVM	Economy	WIA	Import	State of affairs
Nuclear weapon	Development aid	Collaboration	Agriculture	Assessment
Military policy	Assessment	SL	Assessment	Economy
Check	Treaty	Research	International trade	Crude oil
buendnis	Economic policy	Science	Industry	Energy economy
Assessment	Conference	Entrepreneur association	Investment	Situation
Discussion	Financial policy	Economic policy	Cooperation	Research
Missile	Development	Chemical industry	Currency	Nuclear research
Space	Financing	Embargo	Socialist countries	Nuclear industry
Economics & military	**Military technology & strategy**	**Military strategy & international conflicts**	**Enemy activity (foreign nations & intelligence)**	
Payment transactions	Military technology	Military affairs	Enemy activity	
Loan	Armed forces	Armed forces	PID	
Military affairs	Military affairs	Military policy	Contact	
Armed forces	Armament	Conflict settlement	Relation	
Structure	Development	Middle-East policy	Enemy organisation	
Financial concerns	Collaboration	Relation	Organisation	
Staffing	Electronics	Stance	Ostforschung	
schulden	Research	Differences	Idelogical diversion	
Infrastructure	Military policy	Domestic policy	Intelligence agency	
Budget	Air force	Assessment	SL	
Financing	Planning	State of interior affairs	Anti communism	
krisenplanung	Reconnaissance	Liberation movement	Assessment	
Economy	Assessment	War	Conference	
Plan	Land forces	Conflict	Soc/ analysis	
Record	Missile	Arms supply	Counter revolution	

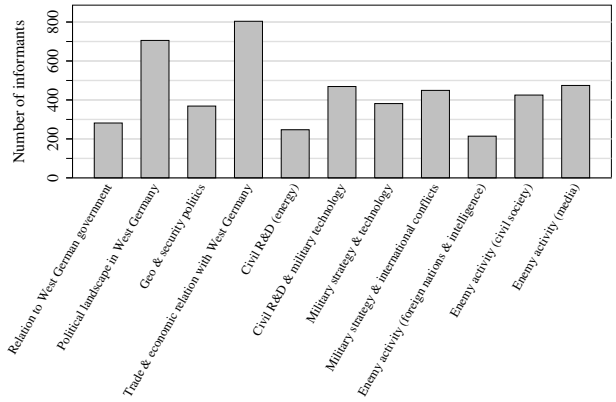
Notes: This table shows the English translation of the 15 most important keywords for each of 14 topics identified by an LDA model with $\alpha = 0.0368421052631579$ and $\eta = 0.04078947368421053$.

Table 3.B.8: Top words per topic, 15 topics

Relation to West German government	**Relation to West Germany**	**Political landscape in West Germany**	**Political & media landscape in West Germany**	**International Relations & Conflicts**
State apparatus	Relation	Party	Election	Domestic policy
Government	Convention	Domestic policy	Party	Relation
Assessment	Traffic	Differences	Media	Differences
Relation	East-West relation	Election	State of interior affairs	Stance
Domestic policy	Status West Berlin	KP	Information technology	Conflict regulation
Journey	Socialist countries	SP	Parliament	Middle East policy
Differences	East policy	Assessment	Election campaign	Opposition
Stance	Stance	State of interior affairs	Assessment	State of interior affairs
East-West relation	Assessment	Economic policy	Program	Assessment
Measure	German question	Union	Press	Liberation movement
Foreign ministry	Journey	Coalition	Coalition	Conflict
Diplomatic representation	Culture	Relation	Staffing	Party
Police	Treaty	Social policy	FF	Difficulty
Talk	Differences	Stance	Public relations	Command
Ministry	Western forces	State of affairs	State parliament	Conference
Geo & Security politics	**Economic relation with West Germany I**	**Economic relation with West Germany II**	**Trade with West Germany**	**Civil R&D**
Security politics	Development policy	Financial concerns	Trade	Research
Disarmament	Payment transactions	Conference	Economy	Industry
East-West relation	Relation	Budget	Economic policy	Economy
Armament	Economy	Economy	East-West relation	Development
Stance	Collaboration	Infrastructure	Commodity	Firm
Negotiation	Trade	Congress	Industry	Collaboration
KSZE	Loan	Law	Export	Science
Conference	Economic policy	Financing	Assessment	IT
Detente	Project	Financial policy	Crude oil	Project
SL	Development aid	Currency	Collaboration	Financing
Party	EL	Committee	Import	Committee
Assessment	Assessment	Law and justice	KIL	Technology
Check	Financial concerns	Report	SL	Support
MVM	Financing	Plan	WIA	Electronics
Nuclear weapon	Investment	Commodity	Firm	Research center
Civil R&D (energy)	**Military technology**	**Military strategy**	**Enemy activity (foreign nations & intelligence)**	**Enemy activity (civil society)**
Energy	Military technology	Military affairs	Enemy activity	Enemy activity
Nuclear energy	Armament	Armed forces	SL	PID
Industry	Military affairs	Military policy	Counter revolution	Organisation
Energy policy	Armed forces	Military strategy	State of interior affairs	Contact
Economic policy	Electronics	Land forces	Enemy organisation	Extremism
State of affairs	Collaboration	Assessment	Intelligence agency	Relation
Energy economy	Reconnaissance	Structure	Influence	Population
Situation	Missile	Exercise	Regime	Left
Assessment	Development	MVM	Ostforschung	Church
Research	Air force	Command	East-West relation	Idelogical diversion
Nuclear research	Planning	War	Armed forces	Maoism
Nuclear industry	ELoKa	Defense	Soc/ analysis	Measure
Power plant	Aircraft	Staffing	Domestic policy	Youth
Reactor	Weapon	Nuclear weapon	Military affairs	Assessment
Tendency	Assessment	Balance of power	Assessment	Conference

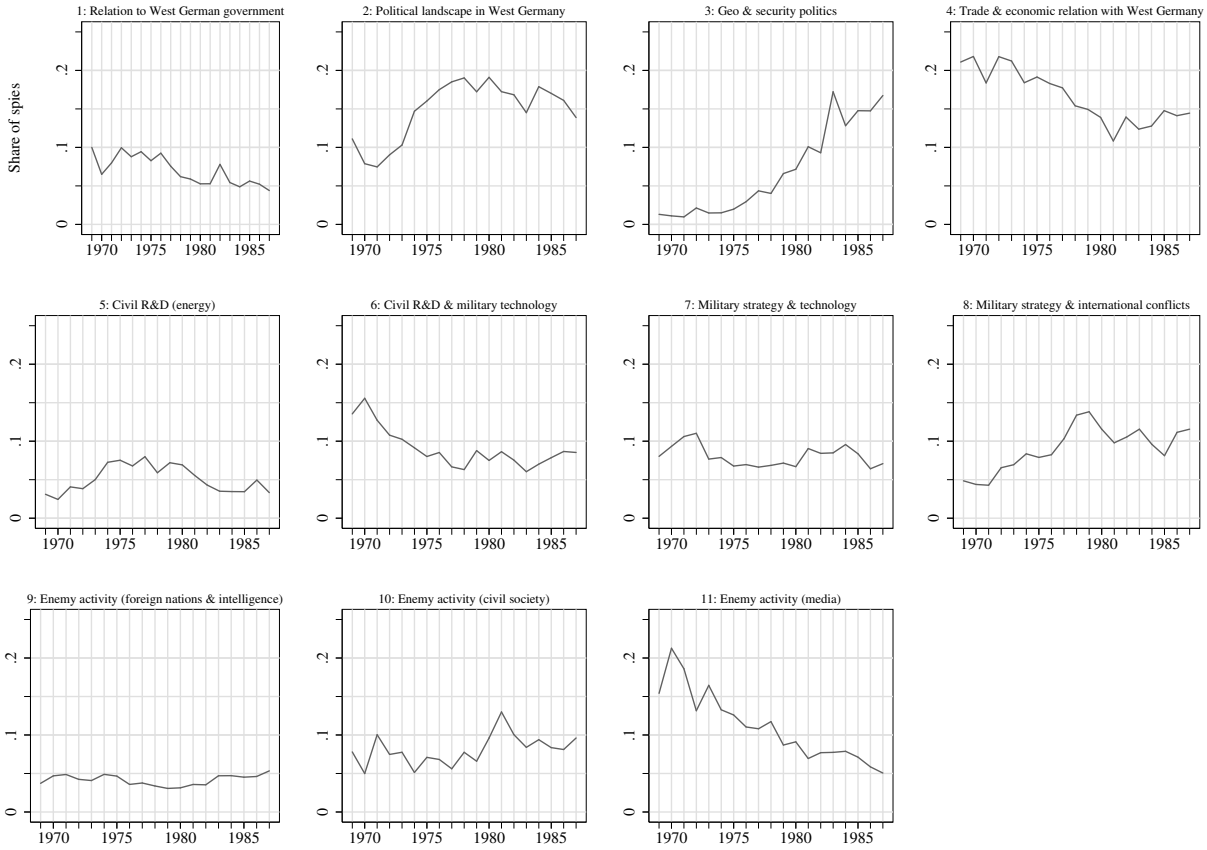
Notes: This table shows the English translation of the 15 most important keywords for each of 11 topics identified by an LDA model with $\alpha = 0.03125$ and $\eta = 0.0328125$.

Figure 3.B.2: Informants by topic



Notes: This figure shows, for each topic, the number of spies which were collecting pieces of information associated with that topic.

Figure 3.B.3: Informants by topic and year



Notes: This figure shows, for each topic and year, the share of spies collecting pieces of information associated with that topic.

Bibliography

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2014). "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration". *Journal of Political Economy 122* (3), 467–506.

Acemoglu, Daron, Tristan Reed, and James A Robinson (2014). "Chiefs: Economic development and elite control of civil society in Sierra Leone". *Journal of Political Economy 122* (2), 319–368.

Adamic, Lada A. and Natalie Glance (2005). "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". *Proceedings of the 3rd International Workshop on Link Discovery*. LinkKDD '05. Chicago, Illinois: Association for Computing Machinery, 36–43.

Akay, Alpaslan, Olivier Bargain, and Klaus F Zimmermann (2017). "Home sweet home? Macroeconomic conditions in home countries and the well-being of migrants". *Journal of Human Resources 52* (2), 351–373.

Alba, Richard D (1990). *Ethnic identity: The transformation of white America*. Yale University Press.

Anderson, Craig A. and Brad J. Bushman (2001). "Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature". *Psychological Science 12* (5), 353–359.

Antecol, Heather, Peter Kuhn, and Stephen J Trejo (2006). "Assimilation via prices or quantities? Sources of immigrant earnings growth in Australia, Canada, and the United States". *Journal of human Resources 41* (4), 821–840.

Aragón, Pablo et al. (2013). "Communication dynamics in twitter during political campaigns: The case of the 2011 Spanish national election". *Policy & internet 5* (2), 183–206.

Baker, Michael and Dwayne Benjamin (1994). "The performance of immigrants in the Canadian labor market". *Journal of labor economics 12* (3), 369–405.

Banerjee, Abhijit et al. (2013). "The diffusion of microfinance". *Science 341* (6144).

Banerjee, Abhijit V (1992). "A simple model of herd behavior". *The quarterly journal of economics 107* (3), 797–817.

Barberá, Pablo (2015). "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". *Political analysis 23* (1), 76–91.

Barberá, Pablo et al. (2015). "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science 26* (10), 1531–1542.

Bauernschuster, Stefan, Oliver Falck, and Ludger Woessmann (2014). "Surfing alone? The Internet and social capital: Evidence from an unforeseeable technological mistake". *Journal of Public Economics 117*, 73–89.

BBSR (2007). "Laufende Raumbeobachtung: Raumabgrenzungen". Tech. rep. Bonn: Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR).

Berry, John W. (2004). "Acculturation". *Encyclopedia of Applied Psychology*. Ed. by Charles D. Spielberger. Elsevier, 27–34.

Berry, John W and Feng Hou (2016). "Immigrant acculturation and wellbeing in Canada." *Canadian Psychology/psychologie canadienne 57* (4), 254.

Besley, Timothy and Robin Burgess (2001). "Political agency, government responsiveness and the role of the media". *European Economic Review 45* (4-6), 629–640.

Bjorvatn, Kjetil et al. (2020). "Teaching through television: Experimental evidence on entrepreneurship education in Tanzania". *Management Science 66* (6), 2308–2325.

Bleakley, Hoyt and Aimee Chin (2004). "Language skills and earnings: Evidence from childhood immigrants". *Review of Economics and statistics 86* (2), 481–496.

Blei, David M and John Lafferty (2009). "Topic Models". *Text Mining: Classification, Clustering, and Applications*. Ed. by Mehran Sahami and Ashok N Srivastava. London: Taylor Francis.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". *Journal of machine Learning research 3* (Jan), 993–1022.

Blume, Lawrence E et al. (1993). "The statistical mechanics of strategic interaction". *Games and economic behavior 5* (3), 387–424.

Boot, Arnout B et al. (2019). "How character limit affects language usage in tweets". *Palgrave Communications 5* (1), 1–13.

Borjas, George J (2015). "The slowdown in the economic assimilation of immigrants: Aging and cohort effects revisited again". *Journal of Human Capital 9* (4), 483–517.

Bracher, Karl Dietrich and Hans-Adolf Jacobsen (2014). Dokumente zur Deutschlandpolitik. Berlin, Germany: Bundesarchiv.

Bramoullé, Yann and Rachel Kranton (2016). "Games Played on Networks". *The Oxford Handbook of the Economics of Networks*. Oxford University Press.

Brekke, Jan-Paul and Grete Brochmann (2015). "Stuck in transit: secondary migration of asylum seekers in Europe, national differences, and the Dublin regulation". *Journal of Refugee Studies 28* (2), 145–162.

"MfS-Handbuch". Ed. by BStU. Berlin, Germany: Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

BStU (2010). *Von Strauß und Schalck-Golodkowski eingefädelt*. URL: https://www.stasi-unterlagen-archiv.de/informationen-zur-stasi/themen/beitrag/von-strauss-und-schalck-golodkowski-eingefaedelt/ (visited on 8/22/2022).

"Der Deutsche Bundestag 1949 bis 1989 in den Akten des Ministeriums für Staatssicherheit (MfS) der DDR" (2013a). Ed. by BStU. Berlin, Germany: Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

BStU (2013b). "Hauptverwaltung A (HV A) Aufgaben - Strukturen - Quellen". Anatomie der Staatssicherheit MfS Handbuch. Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin (2020). "From extreme to mainstream: The erosion of social norms". *American economic review 110* (11), 3522–48.

Bursztyn, Leonardo and Robert Jensen (2015). "How does peer pressure affect educational investments?" *The quarterly journal of economics 130* (3), 1329–1367.

Campante, Filipe, Ruben Durante, and Francesco Sobbrio (2018). "Politics 2.0: The multifaceted effect of broadband internet on political participation". *Journal of the European Economic Association 16* (4), 1094–1136.

Card, David (1995). *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. Ed. by Louis N Christofides, E Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press, 201–222.

Chang, Jonathan et al. (2009). "Reading tea leaves: How humans interpret topic models". *Advances in neural information processing systems 22*.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz (2016). "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment". *American Economic Review 106* (4), 855–902.

Chiswick, Barry R (1978). "The effect of Americanization on the earnings of foreign-born men". *Journal of political Economy 86* (5), 897–921.

——— (1991). "Speaking, reading, and earnings among low-skilled immigrants". *Journal of labor economics 9* (2), 149–170.

Chiswick, Barry R and Paul W Miller (1995). "The endogeneity between language and earnings: International analyses". *Journal of labor economics 13* (2), 246–288.

Chong, Alberto and Eliana La Ferrara (2009). "Television and divorce: Evidence from Brazilian novelas". *Journal of the European Economic Association 7* (2-3), 458–468.

Coleman, James S (1998). *Foundations of social theory*. Harvard university press.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (2014). "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". *Journal of communication 64* (2), 317–332.

Colman, Andrew M (2015). *A dictionary of psychology*. Oxford University Press.

Conley, Timothy G and Christopher R Udry (2010). "Learning about a new technology: Pineapple in Ghana". *American economic review 100* (1), 35–69.

Conover, Michael D et al. (2011). "Predicting the political alignment of twitter users". *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 192–199.

Conover, Michael D et al. (2012). "Partisan asymmetries in online political activity". *EPJ Data science 1* (1), 1–19.

Damm, Anna Piil (2009). "Ethnic enclaves and immigrant labor market outcomes: Quasi-experimental evidence". *Journal of Labor Economics 27* (2), 281–314.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". *Journal of the American society for information science 41* (6), 391–407.

DellaVigna, Stefano, John A List, and Ulrike Malmendier (2012). "Testing for altruism and social pressure in charitable giving". *The quarterly journal of economics 127* (1), 1–56.

DellaVigna, Stefano et al. (2016). "Voting to tell others". *The Review of Economic Studies 84* (1), 143–181.

Dewey, Caitlin (2015). "What you don't know about Internet algorithms is hurting you. (And you probably don't know very much!)" *The Washington Post*.

Diehl, Claudia and Rainer Schnell (2006). ""Reactive ethnicity" or "assimilation"? Statements, arguments, and first empirical evidence for labor migrants in Germany". *International Migration Review 40* (4), 786–816.

Douglas, Paul H (1919). "Is the new immigration more unskilled than the old?" *Quarterly Publications of the American Statistical Association 16* (126), 393–403.

Durlauf, Steven N. and Marcel Fafchamps (2005). "Social Capital". *Handbook of Economic Growth*. Ed. by Philippe Aghion and Steven Durlauf. Elsevier. Chap. 26, 1639–1699.

Dustmann, Christian (1994). "Speaking fluency, writing fluency and earnings of migrants". *Journal of Population economics 7* (2), 133–156.

Dustmann, Christian and Arthur van Soest (2001). "Language fluency and earnings: Estimation with misclassified language indicators". *Review of Economics and Statistics 83* (4), 663–674.

Dümmel, Karsten and Melanie Piepenschneider (2014). *Was war die Stasi? Einblicke in das Ministerium für Staatssicherheit der DDR*. Konrad Adenauer Stiftung.

Easley, David, Jon Kleinberg, et al. (2010). *Networks, crowds, and markets*. Vol. 8. Cambridge university press.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund (2003). "Ethnic enclaves and the economic success of immigrants—Evidence from a natural experiment". *The quarterly journal of economics 118* (1), 329–357.

Eisenbichler, Ernst (2012). *Der Aufsehen erregende Strauß-Deal mit der DDR*. URL: `https://www.br.de/nachricht/inhalt/strauss-kredit-ddr100.html` (visited on 8/22/2022).

Elgie, Robert and Iain McMenamin (2008). "Political fragmentation, fiscal deficits and political institutionalisation". *Public Choice 136* (3), 255–267.

Engelmann, Roger. *MfS-Lexikon*. Stasi Unterlagen Archiv.

Engelmann, Roger, Christian Halbrock, and Frank Joestel (2020). *Vernichtung von Stasi-Akten Eine Untersuchung zu den Verlusten 1989/90*. Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Enikolopov, Ruben et al. (2020). "Social image, networks, and protest participation". SSRN Working Paper No 2940171.

Eslava, Marcela and Oskar Nupia (2010). "Political fragmentation and government spending: Bringing ideological polarization into the picture". SSRN Working Paper No 1554451.

Falck, Oliver, Robert Gold, and Stephan Heblich (2014). "E-lections: Voting Behavior and the Internet". *American Economic Review 104* (7), 2238–65.

Feld, P. Lars et al. (2017). "Fakten statt Stimmungslage Malteser Migrationsbericht 2017". Tech. rep. Köln: Stiftung Malteser.

Franzen, Axel (2003). "Social capital and the Internet: Evidence from Swiss panel data". *Kyklos 56* (3), 341–360.

Friehe, Tim, Markus Pannenberg, and Michael Wedow (2015). "Let bygones be bygones? socialist regimes and personalities in germany". *Socialist Regimes and Personalities in Germany (July 27, 2015)*.

Funk, Patricia (2010). “Social incentives and voter turnout: evidence from the Swiss mail ballot system”. *Journal of the European economic association 8* (5), 1077–1103.

Förster, Günter (1996). “Die Juristische Hochschule des Ministeriums für Staatssicherheit”. Anatomie der Staatssicherheit MfS Handbuch. Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Gaines, Brian J and Jeffery J Mondak (2009). “Typing together? Clustering of ideological types in online social networks”. *Journal of Information Technology & Politics 6* (3-4), 216–231.

Garcia, David et al. (2015). “Ideological and temporal components of network polarization in online political participatory media”. *Policy & internet 7* (1), 46–79.

Gimpel, James G, Frances E Lee, and Joshua Kaminski (2006). “The political geography of campaign contributions in American politics”. *The Journal of Politics 68* (3), 626–639.

Glaeser, Edward L et al. (1992). “Growth in cities”. *Journal of political economy 100* (6), 1126–1152.

Glitz, Albrecht and Erik Meyersson (2020). “Industrial espionage and productivity”. *American Economic Review 110* (4), 1055–1103.

Granovetter, Mark (1978). “Threshold models of collective behavior”. *American journal of sociology 83* (6), 1420–1443.

Green, Alan G and David A Green (1999). “The economic goals of Canada’s immigration policy: Past and present”. *Canadian Public Policy/Analyse de Politiques*, 425–451.

Grenier, Gilles (1984). “The effects of language characteristics on the wages of Hispanic-American males”. *Journal of Human Resources*, 35–52.

Griffiths, Thomas L and Mark Steyvers (2004). “Finding scientific topics”. *Proceedings of the National academy of Sciences 101* (suppl 1), 5228–5235.

Gruzd, Anatoliy and Jeffrey Roy (2014). “Investigating political polarization on Twitter: A Canadian perspective”. *Policy & internet 6* (1), 28–45.

Hahn, Elisabeth et al. (2019). “Predictors of refugee adjustment: The importance of cognitive skills and personality”. *Collabra: Psychology 5* (1).

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.

Hargittai, Eszter, Jason Gallo, and Matthew Kane (2008). “Cross-ideological discussions among conservative and liberal bloggers”. *Public Choice 134* (1-2), 67–86.

Haug, Sonja (2004). “Binationale Ehen und interethnische Partnerschaften in Deutschland: Datenlage und Erklärungsfaktoren”. *Zeitschrift für Familienforschung 16* (3), 305–329.

Hiller, Harry H and Tara M Franz (2004). “New ties, old ties and lost ties: the use of the internet in diaspora”. *New media & society 6* (6), 731–752.

Hoffman, Matthew, Francis Bach, and David Blei (2010). “Online learning for latent dirichlet allocation”. *advances in neural information processing systems 23*.

Hofmann, T (1999). *Probabilistic Latent Semantic Analysis in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Höhne, Jutta et al. (2014). “WSI Report Die Gastarbeiter Geschichte und aktuelle soziale Lage”. Tech. rep. Düsseldorf: Wirtschafts- und Sozialwissenschaftliches Institut.

Jackson, Matthew O. and Yves Zenou (2015). “Games on Networks”. *Handbook of Game Theory with Economic Applications*. Ed. by H. Peyton Young and Shmuel Zamir. Vol. 4. Elsevier. Chap. 3, 95 –163.

Jacob, Marcus and Marcel Tyrell (2010). “The legacy of surveillance: An explanation for social capital erosion and the persistent economic disparity between East and West Germany”. *Available at SSRN 1554604*.

Jensen, Robert and Emily Oster (2009). “The power of TV: Cable television and women's status in India”. *The Quarterly Journal of Economics 124* (3), 1057–1094.

Kama, Amit and Vered Malka (2013). “Identity prosthesis: Roles of homeland media in sustaining native identity”. *Howard Journal of Communications 24* (4), 370–388.

Kasper, Matthias, Christoph Kogler, and Erich Kirchler (2015). “Tax policy and the news: An empirical analysis of taxpayers' perceptions of tax-related media coverage and its impact on tax compliance”. *Journal of behavioral and experimental economics 54*, 58–63.

Kelman, Herbert C (1958). “Compliance, identification, and internalization three processes of attitude change”. *Journal of conflict resolution 2* (1), 51–60.

Kesler, Christel and Irene Bloemraad (2010). “Does immigration erode social capital? The conditional effects of immigration-generated diversity on trust, membership, and participation across 19 countries, 1981–2000”. *Canadian Journal of Political Science/Revue canadienne de science politique 43* (2), 319–347.

Kim, Yonsu and Jae Hong Kim (2021). “What drives variations in public health and social services expenditures? the association between political fragmentation and local expenditure patterns”. *The European Journal of Health Economics*, 1–9.

Kissau, Kathrin (2008). *Das Integrationspotential des Internet für Migranten*. Springer.

Knack, Stephen and Philip Keefer (1997). "Does social capital have an economic payoff? A cross-country investigation". *The Quarterly journal of economics 112* (4), 1251–1288.

Komito, Lee and Jessica Bates (2011). "Migrants' information practices and use of social media in Ireland: networks and community". *Proceedings of the 2011 iConference*, 289–295.

Konopatzky, Stephan (2019). *Dokumentation SIRA - System der Informationsrecherche der Hauptverwaltung A des Ministeriums für Staatssicherheit der DDR*. Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Koser, Khalid (1997). "Social networks and the asylum cycle: The case of Iranians in the Netherlands". *International migration review 31* (3), 591–611.

Koser Akcapar, Sebnem (2010). "Re-thinking migrants' networks and social capital: a case study of Iranians in Turkey". *International migration 48* (2), 161–196.

Kossoudji, Sherrie A (1988). "English language ability and the labor market opportunities of Hispanic and East Asian immigrant men". *Journal of Labor Economics 6* (2), 205–228.

Kosyakova, Yuliya and Herbert Brücker (2020). "Seeking asylum in Germany: Do human and social capital determine the outcome of asylum procedures?" *European Sociological Review 36* (5), 663–683.

Krupp, Hans-Jürgen (2008). "Die Anfänge: Zur Entstehungsgeschichte des SOEP". *Vierteljahrshefte zur Wirtschaftsforschung 77* (3), 15–26.

Kuhn, Peter and Hani Mansour (2014). "Is internet job search still ineffective?" *The Economic Journal 124* (581), 1213–1233.

LaLonde, Robert J, Robert H Topel, et al. (1992). "The assimilation of immigrants in the US labor market". *Immigration and the work force: Economic consequences for the United States and source areas*, 67–92.

Latané, Bibb (1996). "Dynamic social impact: The creation of culture by communication". *Journal of communication 46* (4), 13–25.

Lawrence, Eric, John Sides, and Henry Farrell (2010). "Self-segregation or deliberation? Blog readership, participation, and polarization in American politics". *Perspectives on Politics 8* (1), 141–157.

Lichter, Andreas, Max Löffler, and Sebastian Siegloch (2021). "The long-term costs of government surveillance: Insights from stasi spying in East Germany". *Journal of the European Economic Association 19* (2), 741–789.

Lindner, Sebastian (2011). *Mauerblümchen Kulturabkommen*. URL: https://www.bpb.de/themen/deutschlandarchiv/53911/mauerbluemchen-kulturabkommen/ (visited on 8/22/2022).

Lüdering, Jochen and Peter Tillmann (2020). "Monetary policy on twitter and asset prices: Evidence from computational text analysis". *The North American Journal of Economics and Finance 51*, 100875.

Macrakis, Kristie, Thomas Wegener Friis, and Helmut Müller-Enbergs (2009). *East German foreign intelligence: myth, reality and controversy*. Routledge.

Mang, Constantin (2012). "Online job search and matching quality". Tech. rep. Ifo Working Paper.

Manning, Christopher and Hinrich Schutze (1999). *Foundations of statistical natural language processing*. MIT press.

Marat, Aizhamal (2016). "Uyghur digital diaspora in Kyrgyzstan". *Diaspora Studies 9* (1), 53–63.

McGee, John and Tanya Sammut-Bonnici (2015). "Network Externalities". *Wiley Encyclopedia of Management*. John Wiley Sons, Ltd, 1–5.

McManus, Walter, William Gould, and Finis Welch (1983). "Earnings of Hispanic men: The role of English language proficiency". *Journal of labor Economics 1* (2), 101–130.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001). "Birds of a feather: Homophily in social networks". *Annual review of sociology 27* (1), 415–444.

Melkote, Srinivas R and DJ Liu (2000). "The role of the internet in forging a pluralistic integration: A study of Chinese intellectuals in the United States". *Gazette (Leiden, Netherlands) 62* (6), 495–504.

Meyer, Travis R et al. (2019). "A year in Madrid as described through the analysis of geotagged Twitter data". *Environment and Planning B: Urban Analytics and City Science 46* (9), 1724–1740.

Montalvo, José G and Marta Reynal-Querol (2005). "Ethnic polarization, potential conflict, and civil wars". *American economic review 95* (3), 796–816.

Mosleh, Mohsen et al. (2021). "Shared partisanship dramatically increases social tie formation in a Twitter field experiment". *Proceedings of the National Academy of Sciences 118* (7).

Müller-Enbergs, Helmut (2001). *East German foreign intelligence: myth, reality and controversy*. Ch. Links.

Müller-Enbergs, Helmut. *MfS-Lexikon*. Stasi Unterlagen Archiv.

Müller-Enbergs, Helmut (2007). *Rosenholz - Eine Quellenkritik*. Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Oetzinger, Stephan (2016). *Der Milliardenkredit – Rettete Strauß die DDR?* URL: https://www.csu-geschichte.de/themen/detail/der-milliardenkredit-rettete-strauss-die-ddr/ (visited on 8/22/2022).

Olken, Benjamin A. (2009). "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages". *American Economic Journal: Applied Economics 1* (4), 1–33.

Panagakos, Anastasia (2003). "Downloading new identities: Ethnicity, technology, and media in the global Greek village". *Identities: Global studies in culture and power 10* (2), 201–219.

Perez-Truglia, Ricardo (2018). "Political conformity: Event-study evidence from the United States". *Review of Economics and Statistics 100* (1), 14–28.

Perez-Truglia, Ricardo and Guillermo Cruces (2017). "Partisan interactions: Evidence from a field experiment in the united states". *Journal of Political Economy 125* (4), 1208–1243.

Porter, Matthew and Nick Haslam (2005). "Predisplacement and postdisplacement factors associated with mental health of refugees and internally displaced persons: a meta-analysis". *Jama 294* (5), 602–612.

Putnam, Robert D. (1993). "The Prosperous Community: Social Capital and Public Life". *The American Prospect* (13), 35–42.

——— (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Ricciuti, Roberto (2004). "Political fragmentation and fiscal outcomes". *Public choice 118* (3), 365–388.

Rivera-Batiz, Francisco L (1990). "English language proficiency and the economic progress of immigrants". *Economics Letters 34* (3), 295–300.

——— (1992). "English language proficiency and the earnings of young immigrants in US labor markets". *Review of Policy Research 11* (2), 165–175.

Rupasingha, Anil, Stephan J Goetz, and David Freshwater (2006). "The production of social capital in US counties". *The journal of socio-economics 35* (1), 83–101.

Satyanath, Shanker, Nico Voigtländer, and Hans-Joachim Voth (2017). "Bowling for fascism: Social capital and the rise of the Nazi Party". *Journal of Political Economy 125* (2), 478–526.

Schmole, Angela (2011). "Hauptabteilung VIII Beobachtung, Ermittlung, Durchsuchung, Festnahme". Anatomie der Staatssicherheit MfS Handbuch. Der Bundesbeauftragte

für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU).

Schoeni, Robert F (1997). "New evidence on the economic progress of foreign-born men in the 1970s and 1980s". *Journal of Human Resources*, 683–740.

Selvage, Douglas and Walter Süß (2019). *Staatssicherheit und KSZE-Prozess: MfS zwischen SED und KGB (1972–1989)*. Vol. 54. Vandenhoeck & Ruprecht.

Smither, Robert and Marta Rodriguez-Giegling (1982). "Personality, demographics, and acculturation of Vietnamese and Nicaraguan refugees to the United States". *International Journal of Psychology 17* (1-4), 19–25.

Son, Juyeon (2015). "Immigrant incorporation, technology, and transnationalism among Korean American women". *Journal of International Migration and Integration 16* (2), 377–395.

Statistisches Bundesamt, destatis (2015). "Zeitverwendungserhebung 2012/2013". Tech. rep. Wiesbaden: Statistisches Bundesamt.

Strömberg, David (2004). "Radio's impact on public spending". *The Quarterly Journal of Economics 119* (1), 189–221.

Sunstein, Cass R (2001). *Republic.com*. Princeton University Press.

Tainer, Evelina (1988). "English language proficiency and the determination of earnings among foreign-born men". *Journal of Human Resources*, 108–122.

Tam Cho, Wendy K (2003). "Contagion effects and ethnic contribution networks". *American Journal of Political Science 47* (2), 368–387.

Tang, Jian et al. (2014). "Understanding the limiting factors of topic modeling via posterior contraction analysis". *International Conference on Machine Learning*. PMLR, 190–198.

Valente, Thomas W (2005). "Network models and methods for studying the diffusion of innovations". *Models and methods in social network analysis 28*, 98–116.

Van Alstyne, Marshall and Erik Brynjolfsson (2005). "Global village or cyber-balkans? Modeling and measuring the integration of electronic communities". *Management Science 51* (6), 851–868.

Walther, Lena et al. (2020). "Living conditions and the mental health and well-being of refugees: evidence from a large-scale German survey". *Journal of immigrant and minority health 22* (5), 903–913.

Waters, Mary C (1990). *Ethnic options: Choosing identities in America*. Univ of California Press.

Yardi, Sarita and Danah Boyd (2010). "Dynamic debates: An analysis of group polarization over time on twitter". *Bulletin of science, technology & society 30* (5), 316–327.

Ye, Jiali (2005). “Acculturative stress and use of the Internet among East Asian international students in the United States”. *CyberPsychology & Behavior 8* (2), 154–161.

Yin, Hang (2015). “Chinese-language cyberspace, homeland media and ethnic media: A contested space for being Chinese”. *New Media & Society 17* (4), 556–572.

Yoon, Kyong (2017). “Korean migrants’ use of the internet in Canada”. *Journal of International Migration and Integration 18* (2), 547–562.

Young, H Peyton (2020). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Zagórski, Piotr (2021). “Too much to choose from? The long-term effects of political fragmentation on electoral turnout”. *Politics*.

Zavodny, Madeline (2006). “Does watching television rot your mind? Estimates of the effect on test scores”. *Economics of Education review 25* (5), 565–573.