Ethnic insults in YouTube comments: Social contagion and selection effects during the refugee crisis in Germany

Christoph Spörlein*

Otto-Friedrich-Universität Bamberg

Elmar Schlüter

Justus-Liebig-Universität Giessen

Tables: 0 (+2)

Figures: 5

Total word count: 7.404 (excluding Tables and Figures).

*Corresponding author:
Dr. Christoph Spörlein
Lehrstuhl für Soziologie, insbesondere Sozialstrukturanalyse
Otto-Friedrich-Universität Bamberg
Feldkirchenstraße 21
D-96052 Bamberg
Germany
Phone: +49 (0)951-863-3649
Email: christoph.spoerlein@uni-bamberg.de

Ethnic insults in YouTube comments: Social contagion and selection effects during the refugee crisis in Germany

Abstract: “In this article, we investigate the extent to which ethnic insults propagate through comment networks in YouTube videos from four German political talk shows with the largest audience reach. We argue that comments incorporating ethnic insults signal social norms and embolden others to emulate offensive and abusive behaviour, therefore potentially contributing the contagiousness of insulting commenting. Moreover, periods of highly salient intergroup conflict (i.e., sexual assaults and terrorist attacks), further reduce inhibitions to publicly post insulting content thereby multiplying the spread of this behaviour. To evaluate our claims, we construct a repeated cross-sectional dataset from the comment networks in YouTube videos as well as a pseudo-panel of highly active users which is used to gauge the impact of social selection. Results indicate that the use of ethnic insults in online comment sections appears socially contagious. Specifically, we find that the presence of an ethnic insult increases the prevalence of insulting comments by 8 in the repeated cross-sectional dataset and by 4 percentage points in the user pseudo-panel respectively. The social contagion effects intensify in the aftermath of violent incidents, tripling for the full sample and increasing by 50 percent for the panel of active commenters. Lastly, our empirical findings suggest the social contagiousness of ethnic insults is largely a function of social selection processes.”

1 Introduction

The European refugee crisis generated far-reaching changes in many of the major receiving countries including Germany, France, and Sweden. The large and at times virtually uncontrolled influx of individuals was associated with divergent reactions from politicians and resident populations alike. Responses ranged from building fences in Hungary to Angela Merkel's "we can do this!"; from offering individuals a new home to maliciously torching house about to be occupied by refugees (BBC, 2015; The Guardian, 2015; DW, 2016; The Telegraph, 2016). The near-simultaneous success of right-wing populist movements in European elections is tightly linked to the refugee crisis and can be characterized as a reaction of those population segments who do not agree with the more accommodating forces in European politics (Holmes and Castañeda, 2016; Pisoiu and Ahmed, 2016; Rucht, 2018).

Societal change that some individuals perceive as threatening in general terms may elicit a whole spectrum of reactions and coping strategies – from resigned adaption to actively engaging with the process and trying to counter its impact (Pinquart and Silbereisen, 2004). For a minority, this engagement may mean resorting to violent methods. From the beginning of the refugee crisis (mid-2014) to its highpoint in mid-2016, there has been a roughly 60 percent increase in violent incidents attributed to extreme right-wing individuals or groups in Germany (Destatis, 2018). The vast majority of individuals will however refrain from physical violence and choose other forms of engagement such as political protesting, voicing disconcert with current affairs, or discussing them publicly - for example on social media platforms. Especially on (non-partisan) online platforms, public discussions often derail because individuals tend to be highly emotionally engaged with social changes that impact their daily life and may resort to various forms of verbal attacks targeting individuals or whole minority groups with defamatory slurs (Kwon and Gruzd, 2018; Ottoni et al., 2018). Ethnic insults in online discussions are thus another expression of intergroup conflict[1].

In this article, we study the extent of, and change in, ethnic insults in the comment sections of YouTube videos from the four most popular German political talk shows before, during, and after the height of the European refugee crisis. The talk show setting was chosen because a comparatively large segment of the German population regularly watches the four selected talk shows (i.e., around 15 percent market share [Merkur, 2017]) and they cover a broad range of current topics ranging from content unrelated to ethnic intergroup conflict such as "sugar as a drug", "negative interest rates", or "the outcome of large-scale sporting events" to direct discussions of the refugee crisis and its consequences for life in Germany or Europe, more broadly. In addition, these talk shows strive for non-partisanship and spur considerable public discussion following their airing on television. Episodes of the talk shows are uploaded to YouTube shortly after airing and provide users with the opportunity to comment and engage in online discussions.

We are interested in the process of social contagion whereby some behavior, such as the use of ethnic insults, spreads through a network of users (Christakis and Fowler, 2013). Specifically, we investigate whether prior commenting behavior of actors in online forums creates conditions conducive or suppressive of voicing highly derogatory opinions towards others generally and minorities specifically. Prior research demonstrates the presence of social contagion effects for a variety of behavioral and emotional outcomes such as smoking, divorce and cooperation (see Christakis and Folwer, 2013 for a recent review), suicide (Marsden, 2001), innovations (Rogers, 1995), emotions (Coviello et al., 2014; Kramer, Guillory and Hancock, 2014), prosocial behavior (Fowler and Christakis, 2010; Tsvetkova and Macy, 2014; Lacetera, Marcis and Mele, 2016), attitudes (Howard and Gengler, 2001; Bond, 2018), internet memes (Guadagno et al., 2013; Hodas and Lerman, 2014; Johann and Bülow, 2018) as well as prejudice (Soral, Bilewicz and Winiewski, 2017), interpersonal swearing (Kwon and Gruzd, 2018), interpersonal and xenophobic violence (Fagan, Wilkinson

and Davies, 2007; Braun and Koopmans, 2014) and ethnic hostility (Fox, 2010; Bauer et al., 2018). Our first research questions is as follows: *do preceding social media comments containing insults targeted at ethnic minorities increase the probability that subsequent comments also target ethnic minorities*?

In addition, we contend that intergroup conflict - whether in person or through online comments - is shaped by and responsive to impactful historic incidents, especially when those incidents are related to core dimension of the intergroup conflict (Hainmueller and Hopkins, 2014; Hellwig and Sinno, 2016; Czymara and Schmidt-Catran, 2017; Jäckle and König, 2018). In our study period, three events stand out in particular: the sexual assaults during New Year's Eve 2015 in Cologne and two terrorist attacks: the axe attack in a regional train in Würzburg (July 18, 2016) and the Christmas market attack in Berlin (December 19, 2016). These incidents are central to German debates regarding the appropriate societal response to an influx of culturally dissimilar refugees and should elicit strong reactions from all parts of the discussion spectrum. The increased salience of insulting comments targeted towards ethnic minorities in the aftermath of such incidents can create a normative environment conducive to spreading insulting comments by signaling that insulting commenting behavior is numerically prevalent and virtually goes unpunished due to online anonymity. Our second research question asks *whether ethnic insults are more contagious in the aftermath of important incidents such as mass-scale sexual assault and terrorist attacks*.

In total, we collected data from 90 videos, covering 5,152 comments posted between June 2015 and May 2017. We analyze comment-reply threads in which users post comments and others respond to them. For each comment, we identify whether it contains ethnic insults by using two custom dictionaries containing insulting words and words with ethnic connotations. We analyze the relationship between insults in preceding comments and actor's insulting behavior, as well as how this relationship changes during periods of influential

incidents using multilevel modeling techniques for repeated cross-sectional data. Because commenting in online social networks can be highly selective, we also construct a user pseudo-panel to track the commenting behavior of highly active users over time. Social media comments in some respect represent a more natural situation for researchers to observe interaction patterns where reactions are not elicited by survey questions, not structured or framed by the researcher, and likely considerably less prone to exhibit social desirability biases.

2 Social contagion, selection effects and ethnic insults

Experiencing changes in one's immediate environment due to societal processes that are out of one's control can be highly mobilizing. This is even more so when people perceive environmental changes as direct threats to their identity and cultural values (Jaspers, 1998; van Dyke and Soule, 2002; Almeida, 2003; Ayoub, 2014). Online commenting represents one of many options open for individuals to formulate their grievances with current circumstances. Due to its anonymity, it is oftentimes the option with the least potential for interpersonal conflict with close associates such as friends or family, and also the option with little potential for real-world consequences and sanctions to inappropriate behavior (Álvarez-Benjumea and Winter, 2018). The increasing volume of individuals using online comments to discuss current political affairs generates network structures that set the stage for social contagion processes to occur. Social contagion ideas rest on the notion that attitudes, behavior, or social phenomena spread across a population of connected actors much in the same ways as viral or bacterial pathogens spread through a population (Christakis and Fowler, 2013). Hence, for social contagion processes to take hold, networks of actors need to be connected. Moreover, adjacency to contagious actors increases the likelihood of infection and contagion tends to spread faster across denser networks (Scott, 2000; Centola, 2010).

In the context of online discussion platforms, social contagion processes manifest by "early adopters" using profanity and ethnic insults to attack others (and the opinions they hold) and thereby effectively lowering the threshold for what is deemed acceptable behavior towards other commenters. Hence, observing the past commenting behavior informs individuals about the content of social norms prevalent within each comment thread: repeatedly reading insulting content may reinforce the perception that using this type of language is typical behavior (of a subset) of other commenters (i.e., descriptive social norms; Cialdini and Goldstein, 2004). Moreover, individuals learn to what extent using insulting

language is acceptable by observing how different types of comment interactions are sanctioned by other users or moderators (i.e. injunctive social norms; Bicchieri, 2005). In many cases, what is deemed deviant in real-life interpersonal interactions, is rarely challenged or negatively sanctioned in anonymous online interactions. Obviously, individuals will differ in their susceptibility to "contracting" specific types of behavior just as some individuals' immune systems are easily overwhelmed by certain contagious pathogens whereas others are more resistant. The social contagion mechanism is therefore unlikely to work for all individuals in the same way – especially when it comes to incorporating ethnic insulting behavior in commenting due to strong social norms against it. For instance, it is less plausible that individuals holding highly positive attitudes towards ethnic minorities will suddenly engage in the use of ethnic insults simply by observing descriptive and/or injunctive norms that suggest more or less widespread support of this behavior (Cialdini and Goldstein, 2004; Centola and Macy, 2007). The knowledge of what is *generally* considered deviant behavior does not simply cease to exist when observing deviant behavior in a *particular* social setting (Therborn, 2002). However, observing social norms conducive to ethnic insults can elicit comments with similar content by individuals who hold *a priori* neutral to somewhat negative attitudes towards ethnic minorities but who would normally not voice them because they are well aware of the strong norms against publicly expressing anti-minority sentiment in real-life as well as the consequences of doing so (Ford, 2008; Blinder, Ford and Ivarsaten, 2013). Albeit, the anonymity of online commenting removes the threat of sanctions or consequences and may thereby contribute to lower levels of immunity to contagiousness of ethnic insults (Postmes et al., 2001). Following this line of reasoning, we hypothesize that *the probability of commenters using ethnic insults increases when the previous comments include content considered insulting to ethnic minorities* (H1, social contagion effect).

Another social or contextual condition contributing to the spread of ethnic insults are incidents related to threat perceptions such as extreme forms of intergroup conflict (e.g., terrorist attacks or other forms of violent conflict). These incidents tend to temporarily affect average societal attitude towards ethnic minorities thereby adjusting the injunctive norms towards publicly voicing negative attitudes towards minorities (Das et al., 2009; Hopkins, 2010; Hellwig and Sinno, 2016; Jäckle and König, 2018). Among the few studies investigating the link between terrorist attacks and attitudes towards minorities, Legewie (2013) conducted a quasi-experimental study exploiting the fact that the terrorist attacks in Bali in 2002 and Madrid in 2004 coincided with the fieldwork of two large-scale European cross-national surveys. For both attacks, attitudes became more negative. As a secondary finding, he reported stronger effects for terrorist attacks "closer to home", suggesting stronger reactions to more immediate threats. For the period and national context relevant to this study, Czymara and Schmidt-Catran (2017) demonstrated that the sexual assaults of New Year's Eve in Cologne substantially reduced acceptance of those refugee origin groups connected with the event. More relevant to the approach of our article, namely a focus on behavioral rather than attitudinal changes, Jäckle and König (2018) investigate the relationship between threatening events on violent attacks against refugees in Germany between 2015 and 2016. Mirroring findings in the literature on attitudes towards immigrants, events such as crimes committed by refugees, police raids, and terror attacks in Germany and neighboring countries were associated with a higher incidence of anti-refugee violence. In sum, the negative shift on attitudes towards minorities in the aftermath of attacks in combination with an increased threat perception may make some individuals more susceptible to social contagion effects. Hence, we hypothesize that *social contagion effects will increase in strength during and shortly after major incidents of violent intergroup conflict such as terrorist attacks* (H2, social contagion multiplier effect).

The idea that heightened threat perceptions represent a mobilizing force point to an important and fundamental issue with empirically identifying social contagion effects. Many of the empirical patterns attributed to social contagion effects could have theoretically emerged from selection effects in that individuals tend to associated themselves with similar others (Marsden, 1988; McPherson, Smith-Lovin and Cook, 2001; Lewis, Gonzalez and Kaufman, 2012). In other words, the similarity of attitudes and behaviors exhibited by network members does not necessarily result from mutual exposure to a contagious agent but rather because networks are founded on their prior similarity - that is their stronger threat perception and/or propensity to use ethnic insulting comments. Moreover, selection issues may be more stringent in online (commenting) networks because they differ from real-life social networks in a number of important characteristics: online networks are mainly the result of interactions based on social selection whereas selection mechanisms often play a smaller role in real-world social settings such as schools, churches, associations or the workplace (Putnam, 2000). To participate in online discussions, actors need to make a series of conscious decisions to self-select into situations and context where they interact with others in written form. Interpersonal textual exchanges are thus considerably less spontaneous than real-life encounters where - for the sake of smooth communication - conversations tend to progress much faster, leave less room for editing and structuring responses and are deplete of non-verbal or environmental cues (Okdie et al., 2011; Lapidot-Lefler and Barak, 2012). Writing comments on the other hand requires individuals to engage in structuring one's thought or argument and typing it as coherently as possible without the time pressure of interpersonal communication. The additional investment of cognitive and other resources points to the presence of substantial intrinsic motivation to engage in online commenting - and thus to the presence of strong selection effects. In order to disentangle the contribution of selection effects to empirical patterns of contagious ethnic insulting commenting, we also test social contagion (H1) and social contagion multiplier effects (H2) relying on user pseudo-

panels which enable us to control for selection effects regarding user-constant traits such as the propensity to engage in online discussions of minority-related topics or the propensity to post insulting content.

3. Data and Methods

3.1 Data

To cover the height of the European refugee crisis as well as major incidents in Germany, our data are based on comments posted between June 2015 and May 2017 in the comment sections of YouTube videos for weekly German political talk-shows with the largest audience reach: "*Hart aber Fair*", *Maybrit Illner, Sandra Maischberger* and *Anne Will*. Guests tend to be politicians or representatives from interest groups related to the respective show's topic and typically cover a broad spectrum of the currently prevailing opinions. We categorized each episode as dealing with immigration, refugees, integration, or some other topic. We drew a 25 percent random sample from the episodes categorized as "other topics" to serve as a set of control comments. We then searched YouTube for the show and topic title and recorded the video-URLs for those videos that include the full episode and have at least one comment with a reply (i.e., the smallest possible comment network). The video-URLs are then used to scrape all comments for each video using the SocialMediaLab-package for R (Graham, Ackland and Chan, 2017) as well as respective video metadata using the tuber-package (Sood, Lyons and Muschelli, 2018). In effect, there are three two of YouTube comments: comments with no replies and comment threads consisting of a parent comment where the original poster reacted to the video and any number of child comments nested within the parent comment. For the purpose of studying social contagion, only parent comments and their children are useful. Hence, our data structure is hierarchical with child comments nested in parent comments nested in videos. In total, we collected data from 90 videos with 5,152 comments, 2,681 of which formed a comment network with individuals replying to the original posters' comment.

There are numerous resources to construct (ethnic) insult/swearing dictionaries for English-language text data but unfortunately not for German. In addition, the prevalence of

composite nouns in German creates a comparatively large set of candidate words to be considered as (ethnic) insults. In order to measure instances of ethnic insults, we therefore created two custom dictionaries consisting of stemmed insults identified by manually going through roughly half of all videos (see Appendix Table A1)[2]. One dictionary collects common insults which by themselves do not have an ethnic component (i.e., fuck, shit, ass, dumb). For instance, this comment represents a (real) example of interpersonal insulting behavior (i.e., a user is called by name in combination with the occurrence of an insulting word): "@username du bist krank… solche Idioten wie du braucht keiner Arschloch" [you are sick… nobody needs idiots like you asshole]. Conversely, ethnic insults are identified by the presence of words in the general insults dictionary and words of the second dictionary containing words with explicit ethnic connotations (i.e., arab, goat, muslim, gypsy): "@username träum weiter, moslem. Ganz Deutschland hasst deinen beschissenen Islam. Deine Drecksreligion ist erledigt" [dream on, muslim. The whole of Germany hates your fucking islam. Your dirty religion is finished]. Comments are then categorized as containing ethnic insults by recording whether at least one insulting word and at least one word with ethnic connotations are mentioned by matching the stemmed words of the dictionaries with the comment strings.

Main independent measures

In order to test the hypothesized social contagion and social contagion multiplier effects, we record for each comment whether the *preceding comment* contains ethnic insults or not. For the first child comment, the parent value is used. The second central measure records whether or not one of three *events* (New Year's Eve 2015 and the two terrorist attacks in Würzburg and Berlin) occurred within a four-week window before the comment's post date. For example, comments up to but not including January 2016 score 0 whereas comments posted in January 2016 score 1 to indicate the aftermath of an impactful event (in this case New Year's Eve in Cologne). This time frame was chosen because talk shows often take between

one or two weeks to dedicate one or more shows to these attacks and because the political, criminal, and social coming to terms with what has occurred determines public discussions and news cycles for quite some time afterwards[3].

Thread-level controls

In contrast to child comments, threads are not portrayed in chronological order under YouTube videos. We thus include a measure of *thread popularity* to account for the increased visibility of threads with many "I like this" votes and their higher likelihood of receiving additional comments. In addition, we record whether a thread's *parent comment contains ethnic insults* because those comments represent strong signals of prevalent social norms and thereby set the tone of the child comments, potentially increasing the level of overall ethnic insults. And lastly, we include a measure of the *number of child comments* each thread contains because longer threads have a higher chance of starting and containing contagion processes in the first place.

Comment-level controls

In a similar vein, we include a child comment's *word count* to control for differences in the potential to insult between shorter and longer comments. In addition, the *time between the posting of the video in question and a comment* (in days) is controlled for.

Video-level controls

And finally, we include three video-level controls: Insulting may be more prevalent for videos with a *higher percentage of dislikes* as well as for those videos which received on aggregate more polarized user feedback. *Polarization* is measured using the Simpson diversity index which records 1 minus the sum of the squared proportions of likes and dislikes each video received. Theoretical values range from 0 "no polarization", to 0.5 "complete polarization".

As stated above, videos are also grouped into two topic categories: videos with *refugee or related topics* and those without.

All continuous variables are standardized. Descriptive statistics can be found in Appendix Table A2.

3.2 Methods

One advantage of online social network data rests with its versatility regarding potential analytical strategies. As soon as they are posted, videos accumulate comments over time thus resembling repeated cross-sectional dataset or pseudo-panels. And while comments are continuously posted under those videos, we can exploit changes in contextual, time-variant characteristics such as - in our case – the occurrence of incidence likely to shape threat perceptions of minorities. Figure 1 presents a visual representation of our research design for the first part of this study: the three arrows of Figure 1 depict the progression through time of three different videos (A, B and C arrows). The grey bar signifies the occurrence of a major incident and its aftermath. For the purpose of this article, video A and (more so) video B represent interesting test cases to estimate the effect of incidents on changes in commenting content because they both cover varying proportions of time periods with and without incidents. Video B, for example, was posted in a period without a major incident. The change in the incident state takes place around t10 with the following ten time periods being characterized by the incident and its aftermath. Video C can be seen as control case where all commenting is made during time contexts without major incidents.

[Figure 1 about here]

In more traditional comparative research settings, this design is analogous to having time series data on the country-level in cross-national research and combining it with repeated

cross-sectional survey data of individuals to assess the changes in individual behavior associated with the longitudinal changes on the country-time level (Fairbrother, 2013; Spörlein, Schlüter and van Tubergen, 2014; Spörlein, Mouw and Martinez-Schuldt, 2016). We therefore also rely on multilevel methods for repeated-cross-sectional data. Doing so requires the introduction of a video-time level representing the cross-classification of videos and time (measured in weeks since June 1st 2015). Our time-variant measure for violent incidents is subsequently group-mean centered (i.e., centered "within" videos) with the group-mean representing the (time-invariant) cross-sectional component and the de-meaned values the (time-varying) longitudinal component. The resulting cross-sectional and longitudinal components are uncorrelated due to the group-mean centering thus enabling us to estimate their effects simultaneously. In short, the cross-sectional effect indicates whether the comment content systematically differs between videos for which commenting predominantly occurred during incidents (i.e., video B) compared to videos where most commenting occurred in periods without incidents (i.e., video A and more so video C). The – for our study – more important longitudinal effect indicates the change in the content of comments *within* videos due to incident occurring.

The first part of this study relies on linear probability models to estimate differences in the probability of comments containing ethnic insults. More specifically, we estimate multilevel models where child comments are nested in parent comments, videos and the cross-classification of video and time. Note that whereas multiple variables are measured on the child-, parent-, and video-level, only the longitudinal component of incidents is located on the video-time-level (see Appendix Table A2 for more detail). The second part of this study relies on a synthetic user panel to gauge the impact of network selection on the use of ethnic insults. In total, 251 users engaged in commenting behavior in at least two separate videos. Overall, this subsample includes 1,275 comments with single users posting an average of

around 5 comments. OLS-Regressions with user fixed-effects enable us to estimate *within-person changes* in commenting behavior due to incidents. The inclusion of user fixed-effects estimates this effect holding constant effectively time-invariant user-specific attributes such as the propensity to engage in online commenting, propensity to use ethnic insults or latent attitudes towards minorities. This is not to say that attitudes towards minorities are unchangeable individual attributes, the theoretical ideas presented in this article hinge on the idea that they are in fact not fixed. But they are fixed in the sense that extreme shifts are exceedingly rare. Individuals rather tend to change incrementally within the relative narrow confines of their "attitudinal neighbourhood" (e.g., Tuschman, 2013; Lancee and Sarrasin, 2015).

4 Results

We start by presenting the overall time trends for talk show videos with and without refugee topics by plotting the percentage of comments containing ethnic insults. For reference, the three incidents are plotted as grey bars. On average, around 8 percent of the videos included in our sample contain comments with ethnic insults. Figure 2 illustrates ethnic insults are more prevalent in videos of talk shows covering refugee related topics (9 vs 2 percent) suggested highly emotional salience with considerable more heated debates and derogatory attacks aimed at minorities. Descriptively, the incidence of ethnic insults tend to increase with two of the three incidents depicted here: both after the sexual assaults in Cologne and the Axe attack in Würzburg sees an increase in ethnic insults. However, it should be noted that the rate of ethnic insults is on a secular upward trend since the New Year's Eve incident in Cologne. The attack at the Berlin Christmas market even coincides with the overall peak rather than fostering an upward trend. As another cautionary side note, many smaller incidents happened all over 2016 that may have been impactful in shaping local attitudes more so than aggregate national attitudes but YouTube videos are not ideally suited to exploit regional variation within countries.

[Figure 2 about here]

Next, we extend the descriptive analyses using multilevel models for repeated cross-sectional data. Figure 3 plots each variables coefficients (black dot) and associated standard errors (red line). In line with the descriptive results presented in Figure 2, the results from our multilevel models indicate that videos of talk shows with refugee topics elicit more comments with ethnic insults, by around 6 percentage points, relative to videos from shows with other topics (0.06, $p<0.05$). In addition, comments following parent threads that contain content with ethnic insults are also considerably more likely to resort to insulting comments targeted at minorities (0.06, $p<0.05$) suggesting strong "tone-setting" effects by early commenters. The

strongest relationship among the control variables with the probability of posting content with ethnic insults is related the word count of child comments: writing an additional 70 words (~one standard deviation) corresponds to a 7 percentage point increase in posting ethnic insults. None of the remaining control variables appear to be systematically related to the likelihood of ethnic insults. With respect to the theoretically relevant variables, the results provide strong indication for social contagion effects with the likelihood of ethnic insults substantially increasing (+8 percentage points; 0.08, $p<0.05$) when the immediately preceding comment includes ethnic insults - over and above the presence of ethnic insults in the comment networks parent comment. This is also the strongest relationship in the model. Consistent with the overall findings depicted in Figure 2, we do not find support for the idea that ethnic insults *generally* increases in periods with major incidents (i.e., the cross-sectional component; -0.01, $p=0.63$) nor during state changes, that as when incidents occur and shortly thereafter (i.e., the longitudinal component; 0.02, $p=0.19$).

[Figure 3 about here]

Figure 4 present similar findings based on the user pseudo-panel data. Findings regarding ethnic insults are very similar to those based on the full sample however the smaller sample size is associated with a loss of statistical power. Measures indicating if a video covers refugee topics or if the parent comment contains ethnic insults turn insignificant despite their coefficients remaining similar to the estimates on the full sample (0.06, $p=0.15$ and 0.05, $p=0.06$). Word count has the strongest association with the probability of posting ethnic insulting comments (0.08, $p<0.05$). In line with the ideas formulated regarding selection effects, the social contagion effect is considerably reduced compared to the non-panel results with derogatory content in preceding comments increasing the likelihood of ethnic insults in child comments by four percentage points (0.04, $p<0.05$ vs 0.08 in the full sample; see Figure 3). Frequent commenters therefore appear to be substantially less receptive to social contagion

effects. Figure 4 also shows that incidents are again unrelated to the incidence of ethnic insults in this panel of highly active users (0.02, p=0.47). It should be noted again that the goal of this article is not to demonstrate *absolute increases* in the incidence of ethnic insults in relation to incidents of intergroup violent behavior, but rather to investigate the *conditions that facilitate the spread* of insulting commenting.

[Figure 4 about here]

Figure 5 provides evidence for the idea that violent intergroup incidents are associated with a faster spread of socially contagious use of ethnic insults. For the full sample as well as the subsample of frequent commenters, the interaction between social contagion indicator (i.e., whether the immediately preceding comment contains ethnic insults) and the longitudinal measure of incidents occurring are presented. Focusing on the left panel, the social contagion effect of insults in preceding comments almost *triple* during periods of major incidents (0.16, $p<0.05$; 0.16+0.08=24 percentage points) compared to periods without incidents (+8 percentage points, $p<0.05$) suggesting strong multiplier effects. Put differently, in periods without incidents, every 13^{th} comment (100/8=12.5) is contagious whereas in periods with incident, every 4^{th} comment (100/24=4.2) manages to elicit ethnic insults from commenters. Among the frequent commenters in the pseudo-panel (right panel), the multiplier effect is empirically also present but substantively less pronounced. The baseline social contagion effect (4 percentage points) increased by 2 percentage points (0.02, $p<0.05$) in the aftermath of violent incidents (contagiousness of every 25^{th} vs every 17^{th} comment).

[Figure 5 about here]

5 Summary and Conclusion

In this article, we relied on behavioral data on the form of online comments posted in comment threads of YouTube videos from four German talk shows with the largest audience reach. Relying on ideas from the social norms literature, we argued that the use of ethnic insults in online forums can be socially contagious and propagate through online commenting networks (i.e., social contagion effect). Moreover, we contested that specific contextual conditions – namely the occurrence of violent intergroup incidents – and the resulting increased threat perception can make ethnic insults more socially contagious than in periods with little to no intergroup conflict (i.e., social contagion multiplier effects). Our results present evidence for both effects – both in the full, unrestricted sample of comment networks as well as in pseudo-panel data of highly active commenters.

Our findings highlight several important conclusions: first, social contagion effects are found in both samples. Ethnic insults in preceding comments were associated with an 8 and 4 percentage point increase in subsequent commenters mimicking the insulting content. Second, we also find social contagion multiplier effects where ethnic insulting commenting became more viral – in the literal sense – in periods of violent attacks committed by minority members. In the full sample, the contagiousness of insulting commenting behavior almost tripled while it increased by roughly 50 percent among active commenters. And third, comparing the findings from the full sample and the active commenters suggests that the tripling of the social contagion effect during periods of violent incidents is likely largely driven by strong selection effects. That is, there is a tendency of individuals with a higher propensity to employ ethnic insults to seek out videos with refugee-related content. Still, although the empirical patterns suggesting social contagion are largely accounted for by social selection, we still find evidence that violent incidents committed by minority members are associated with an increased willingness to violate norms against derogatory comments

targeting minorities among the highly active segment of commenters - where high activity is interpreted as an a priori high engagement with minority related topics.

Extensions of this study might tackle some of the limitations of this study, primarily providing an empirical account of the social selection mechanism at play. For instance, one extension might investigate the contribution of local, regional and/or national elections as a mobilizing force contributing to social selection effects. Especially parties campaigning on safety-, or immigration-related issues which saw a considerable rise concurrent to the refugee crisis might lead to increasing comment volume and increases in insulting content (Howard and Kollanyi, 2016). In addition, cross-platform sharing of YouTube videos in social networks such as Facebook or Twitter might be another factor contributing to commenter recruitment and thus social selection. Moreover, pursuing a cross-national comparative research agenda could provide valuable insights into the generalizability of the findings as well as the approaches to measuring insulting content across distinct languages. A number of other important European receiving nations for refugees also experienced widely publicized violent attacks or sexual assaults by minority members within the same time frame (e.g., the Paris terror attacks in November 2015 or the London Bridge attack in June 2017) enabling a direct replication of this study.

These findings thus have broader implications for policy responses aimed at promoting more respectful forms of online interaction (ECRI, 2016) and contribute to a growing literature on combating abusive online content. Censoring may be an effective response against the contagiousness of insulting content because removing the contagious agent, essentials prevents or impedes social contagion processes. To be sure, a recent experimental study demonstrated that censoring in the form of deleting abusive comments was more effective than milder interventions such as verbally sanctioning abusive commenters (i.e., counter-speaking; Álvarez-Benjumea and Winter, 2018). However, given

that we found social contagion effects to be considerably superseded by social selection effects, censoring to inhibit the propagation of hateful content may not represent the optimal response because of the potential societal costs regarding curtailing of foundational social values such as freedom of expression (Foxman and Wolf, 2013). Designing effective responses to social selection effects then entails primarily dealing with individual's attitudes towards outgroups generally and minorities specifically and/or individual's propensity to express oneself relying on abusive content. As such, interventions aimed at reducing abusive externalizing behaviors must integrate (social-) psychological insights and address their individual and contextual determinants (Meuleman, Davidov and Billiet, 2009; Paluck and Green, 2009).

Literature

Almeida, P.D. (2003). Opportunity Organizations and Threat-Induced Contention: Protest Waves in Authoritarian Settings. *American Journal of Sociology,* **109**, 345-400.

Álvarez-Benjumea, A. and Winter F. (2018). Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review*, **34**, 223-237.

Ayoub, P.M. (2014). With Arms Wide Shut: Threat Perception, Norm Reception, and Mobilized Resistance to LGBT Right. *Journal of Human Rights*, **13**, 337-362.

Bauer, M., Cahliková, J., Chytilová, J. and Želinský, T. (2018). Social contagion of ethnic hostility. *Proceedings of the National Academy of Science,* **115**, 4881-4886.

BBC. (2015). Migrant crisis: Hungary's closed border leaves many stranded. https://www.bbc.com/news/world-europe-34260071. Retrieved 21.06.2018.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Bond, R.M. (2018). Contagion in social attitudes about prejudice. *Social Influence*, **13**, 104-116.

Braun, R. and Koopmans, R. (2014). Watch the crowd: bystander responses, trickle-down politics and xenophobic mobilization. *Comparative Political Studies,* **47**, 631-658.

Centola, D. and Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, **113**, 702-734.

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, **329**, 1194-1197.

Christakis, N.A. and Fowler, J.H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine,* **32**, 556-577.

Cialdini, R.B. and Goldstein, N.J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology,* **55**, 591-621.

Coviello, L., Sohn, Y., Kramer, A.D.I, Marlow, C., Franceschetti, M., Christakis, N.A. and Fowler, J.H. (2014). Detecting Emotional Contagion in Massive Social Networks. *PLoS One,* **9**, e90315.

Czymara, C.S. and Schmidt-Catran, A.W. (2017). Refugees Unwelcome? Changes in the Public Acceptance of Immigrants and Refugees in Germany in the Course of Europe's `Immigration Crisis´. *European Sociological Review,* **33**, 735-751.

Das, E., Bushman, B.J., Bezemer, M.D., Kerkhof, P. and Vermeulen, I.E. (2009). How terrorism news reports increase prejudice against outgroups: A terror management account. *Journal of Experimental Social Psychology*, **45**, 453-459.

Destatis. (2018). *Rechte Gewalt in Deutschland.* https://de.statista.com/infografik/12021/rechte-gewalt-in-deutschland/. Retrieved: 18.06.2018.

DW. (2016). Arsonists increasingly target refugee shelters in Germany. http://www.dw.com/en/arsonists-increasingly-target-refugee-shelters-in-germany/a-19545693. Retrieved 21.06.2018.

ECRI. (2016). *Recommendation No. 15 on Combating Hate Speech*. Strasbourg: Council of Europe.

Fagan, J., D.L. Wilkinson and Davies, G. (2007). Chapter 36 – Social Contagion of Violence. In: Flannery, D.j, A.T. Vazsonyi and I.D. Waldman (eds.): *The Cambridge Handbook of Violent Behavior and Aggression*. Cambridge: Cambridge University Press.

Fairbrother, M. (2013). Two Multilevel Modeling Techniques for Analyzing Comparative Longitudinal Survey Datasets. *Political Science Research and Methods*, **2**, 119-140.

Ford, R. (2008). Is racial prejudice declining in Britain? *The British Journal of Sociology*, **59**, 609-636.

Fowler, J.H. and Christakis, N.A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, **107**, 5334-5338.

Fox, J. (2010). Is Ethnoreligious Conflict a Contagious Disease? *Studies in Conflict & Terrorism*, **27**, 89-106.

Foxman, A.H. and Wolf, C. (2013). *Viral Hate: Containing Its Spread on the Internet*. New York: Palgrave Macmillan.

Gaurav, S., Lyons, K. and Muschelli, J. (2018). *tuber: Access YouTube from R*. R package version 0.9.7.9001.

Graham, T., Ackland, R. and Chan, C. (2017). *SocialMediaLab: Tools for Collecting Social Media Data and Generating Networks for Analysis*. R package version 0.23.3.

Guadagno, R.E., Rempala, D.M., Murphy, S. and Okdie, B.M. (2013). What makes a video go viral? An analysis of emotional contagion and Internet memes. *Computers in Human Behavior*, **29**, 2312-2319.

Hainmueller, J. and Hopkins, D.J. (2014). Public Attitudes Towards Immigration. *Annual Review of Political Science*, **17**, 225-249.

Hellwig, T. and Sinno, A. (2016). Different groups, different threats: public attitudes towards immigrants. *Journal of Ethnic and Migration Studies*, **43**, 339-358.

Hodas, N.O. and Lerman, K. (2014). The Simple Rules of Social Contagion. *Scientific Reports*, **4**, 4343-4350.

Holmes S.M. and Castañeda, H. (2016). Representing the "European refugee crisis" in Germany and beyond: Deservingness and differences, life and death. *American Ethnologist*, **43**, 12-24.

Hopkins, D.J. (2010). Politicized places: explaining where and when immigrants provoke local opposition. *American Political Science Review*, **104**, 40-60.

Howard, P.N. and Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. arXiv:1606.06356.

Jäckle, S. and König, P.D. (2018). Threatening Events and Anti-Refugee Violence: An Empirical Analysis in the Wake of the Refugee Crisis during the Years 2015 and 2016 in Germany. *European Sociological Review,* 1-16.

Jasper, J.M. (1998). The Emotions of Protest: Affective and Reactive Emotions In and Around Social Movements. *Sociological Forum,* **13**, 397-424.

Johann, M. and Bülow, L. (2018). Die Verbreitung von Internet-Memes. Empirische Befunde zur Diffusion von Bild-Sprache-Texten in den sozialen Medien. In: Fischer, G. and L. Grünewald-Schukalla (eds.*): Originalität und Viralität von (Internet-)Memes. Sonderausgabe von kommunikation@gesellschaft* 19.

Kramer, A.D.I., Guillory, J.E. and Hancock, J.T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences,* **111**, 10780-10783.

Kwon, H.K and Gruzd, A. (2018). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research,* **27**, 991-1010.

Lacetera, N., Macis, M. and Mele, A. (2016). Viral Altruism? Charitable Giving and Social Contagion in Online Networks. *Sociological Science,* DOI 10.15195/v3.a11.

Lancee, B. and Sarrasin, O. (2015). Educated Preferences or Selection Effects? A Longitudinal Analysis of the Impact of Educational Attainment on Attitudes Towards Immigrants. *European Sociological Review*, **31**, 490-501.

Lapidot-Lefler, N. and Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, **28**, 434-443.

Legewie, J. (2013). Terrorist Events and Attitudes toward Immigrants: A Natural Experiment. *American Sociological Review*, **118**, 1199-1245.

Lewis, K., Gonzalez, M. and Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, **109**, 68-72.

Marsden, P. (1988). Homogeneity in confiding relations. *Social Networks,* **10**, 57-76.

Marsden, P. (2001). Is Suicide Contagious? A Case Study in Applied Memetics. *Journal of Memetics – Evolutionary Models of Information Transmission,* **5**.

McMahon, J.M. and Kahn, K.B. (2017). When Sexism Leads to Racism: Threat, Protecting Women, and Racial Bias. *Sex Roles,* **78**, 591-605.

McPherson, Smith-Lovin, M, L. and Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology,* **27**, 415-444.

Merkur. (2017). *Große Quoten-Analyse: Das waren die besten Talkshows 2017.* https://www.merkur.de/tv/grosse-quoten-analyse-waren-besten-talkshows-2017-9463415.html. Retrieved 18.06.2018.

Meuleman, B., Davidov, E. and Billiet, J. (2009). Changing attitudes toward immigration in European societies, 2002-2007: a dynamic group conflict theory approach. *Social Science Research,* **38**, 352-365.

Mullen B. and Smyth, J.M. (2004). Immigrant suicide rates as a function of ethnophaulisms: hate speech predicts death. *Psychosomatic Medicine,* **66**, 343-348.

Okddie, B.M., Guadagno, R.E., Bernieri, F.J., Geers, A.L. and McIarney-Vesotski, A.R. (2011). Getting to know you: Face-to-face versus online interactions. *Computers and Human Behavior,* **27**, 153-159.

Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira Jr., W. and Almeida, V. (2018). Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. arXiv:1804.04096v1.

Paluck, E.L. and Green, D.P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology*, **60**, 339-367.

Pinquart, M. and Silbereisen, R.K. (2004). Human development in times of social change: Theoretical consideration and research needs. *International Journal of Behavioral Development,* **28**, 289-298.

Pisoiu, D. and Ahmed, R. (2016). Capitalizing on Fear: The Rise of Right-Wing Populist Movements in Western Europe. In: Institute of Peace Research and Security Policy at the University of Hamburg. *OSCE Yearbook 2015*. Baden-Baden: Nomos.

Postmes, T., Spears, R., Sakhel, K. and de Groot, D. (2001) Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior. *Personality and Social Psychology Bulletin*, **27**, 1243-1254.

Putnam, R.D. (2000). *Bowling Alone: The Collapse and Revival of American Community.* New York: Simon & Schuster Paperbacks.

Rambaran, A.J., Dijkstra, J.K. and Stark, T.H. (2013). Status-Based Influence Processes: The Role of Norm Salience in Contagion of Adolescent Risk Attitudes. *Journal of Research on Adolescence,* **23**, 574-585.

Rogers, E.M. (1995). *Diffusion of Innovations*. New York: The Free Press.

Rucht, D. (2018). Mobilization Against Refugees and Asylum Seekers in Germany: A Social Movement Perspective. In: Rosenberger, S.; V. Stern and N. Merhaut. *Protest Movements in Asylum and Deportation*, pp. 225-245. Berlin: Springer.

Scott, J. (2000). *Social Network Analysis: A Handbook*. London: Sage Publications.

Soral, W., Bilewicz, M. and Winiewski, M. (2017). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior,* **44**, 136-146.

Spitzberg, B.H. (2014). Toward A Model of Meme Diffusion. *Communication Theory,* **24**, 311-339.

Spörlein, C., van Tubergen, F. and Schlueter, E. (2014). Ethnic intermarriage in longitudinal perspective: Testing structural and cultural explanations in the United States, 1880-2011. *Social Science Research,* **43**, 1-15.

Spörlein, C., Mouw, T. and Martinez-Schuldt, R. (2016). The interplay of spatial diffusion and marital assimilation of Mexicans in the United States, 1980-2011. *Journal of Ethnic and Migration Studies,* **43**, 475-494.

The Guardian. (2015). Germany's response to the refugee crisis is admirable. But I fear it cannot last. https://www.theguardian.com/commentisfree/2015/sep/06/germany-refugee-crisis-syrian. Retrieved 21.06.2018

The Telegraph. (2016). Merkel defiant over refugee policy: "we can do it". https://www.telegraph.co.uk/news/2016/08/31/merkel-defiant-over-refugee-policy-we-can-do-it/. Retrieved 21.06.2018.

Therborn, G. (2002). Back to Norms! On the Scope and Dynmics of Norms and Normative Action. *Current Sociology*, **50**, 863-880.

Tsvetkova, M. and Macy, M.W. (2014). The Social Contagion of Generosity. *PLoS ONE*, **9**: e87275.

Tuschman, A. (2013). *Our Political Nature: The Evolutionary Origins of What Divides Us*. Amherst: Prometheus Books.

van Dyke, N. and Soule, S.A. (2002). Structural Social Change and the Mobilizing Effect of Threat: Explaining Levels of Patriot and Militia Organizing in the United States. *Social Problems*, **49**, 497-520.

Endnotes:

1 See Muller and Smyth (2004) for evidence that ethnic insults are associated with higher suicide rates among immigrants.

2 To ensure that we did not miss important words, we also manually went through a 50 percent sample of comments that our approach indicated did not include ethnic insults.

3 We tested a number of alternative specifications: the one-, two-, and three-week window after the incident date scores a one. All approaches lead to with the same conclusions results with coefficients ranging from 0.28 ($p<.05$) for the one-week window to 0.17 ($p<0.05$) for the three week window.

Author Biography:

Christoph Spörlein is a postdoctoral researcher at the University of Bamberg. His research focuses on question regarding (selective) migration and question regarding immigrant integration in terms of language acquisition, educational attainment, labor and housing market integration as well as interethnic marriage. He recently published in *Research in Social Stratification and Mobility*, *International Migration Review*, *PLOS ONE* and the *Journal of Ethnic and Migration Studies*.

Elmar Schlueter is professor of sociology at the Justus Liebig University, Giessen. His main research interests focus on interethnic relations and immigrant integration, often coupled with methodological applications of multilevel and structural equation modeling.

Figures

Figure 1: Research design

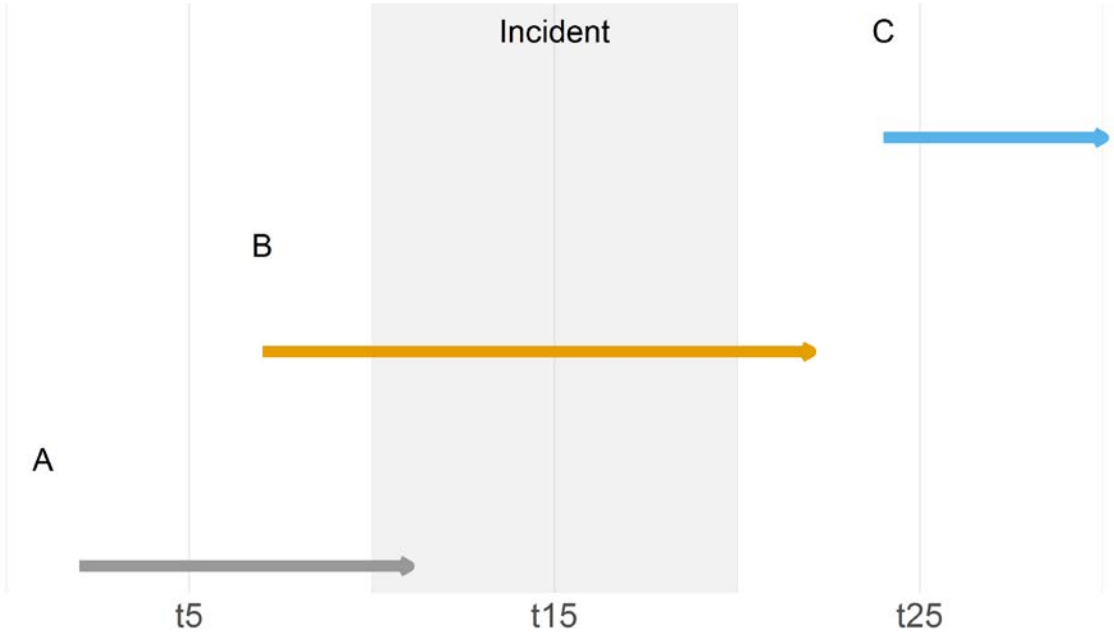


Figure 2: Time trends in ethnic insults, separately for YouTube videos of talk show episodes with and without refugee topics

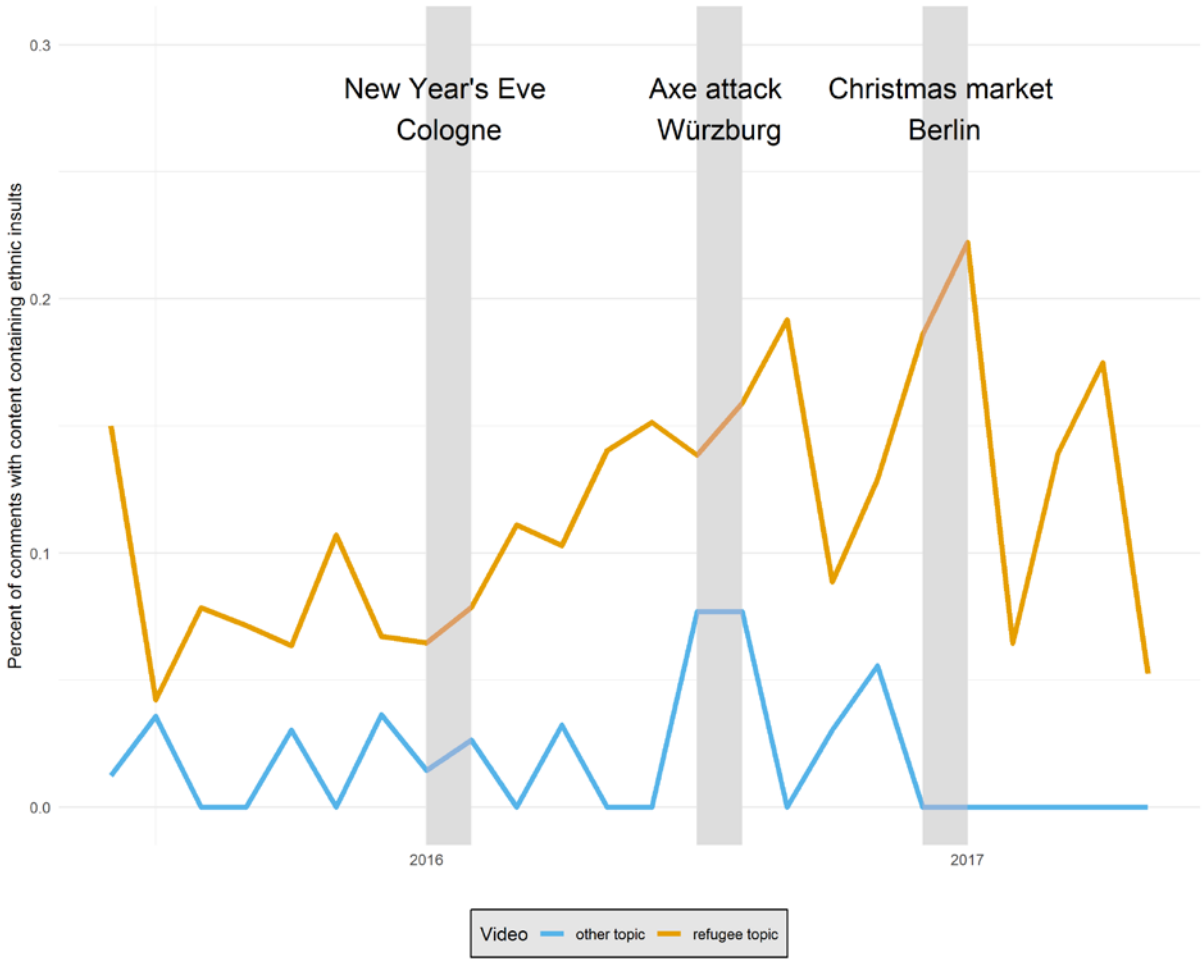


Figure 3: Ethnic insults in YouTube videos of German political talk shows (June 2015-April 2017), linear probability multilevel model, nvideos=90, nvideoxtime=293, nparents=688, ncomments=2,681.

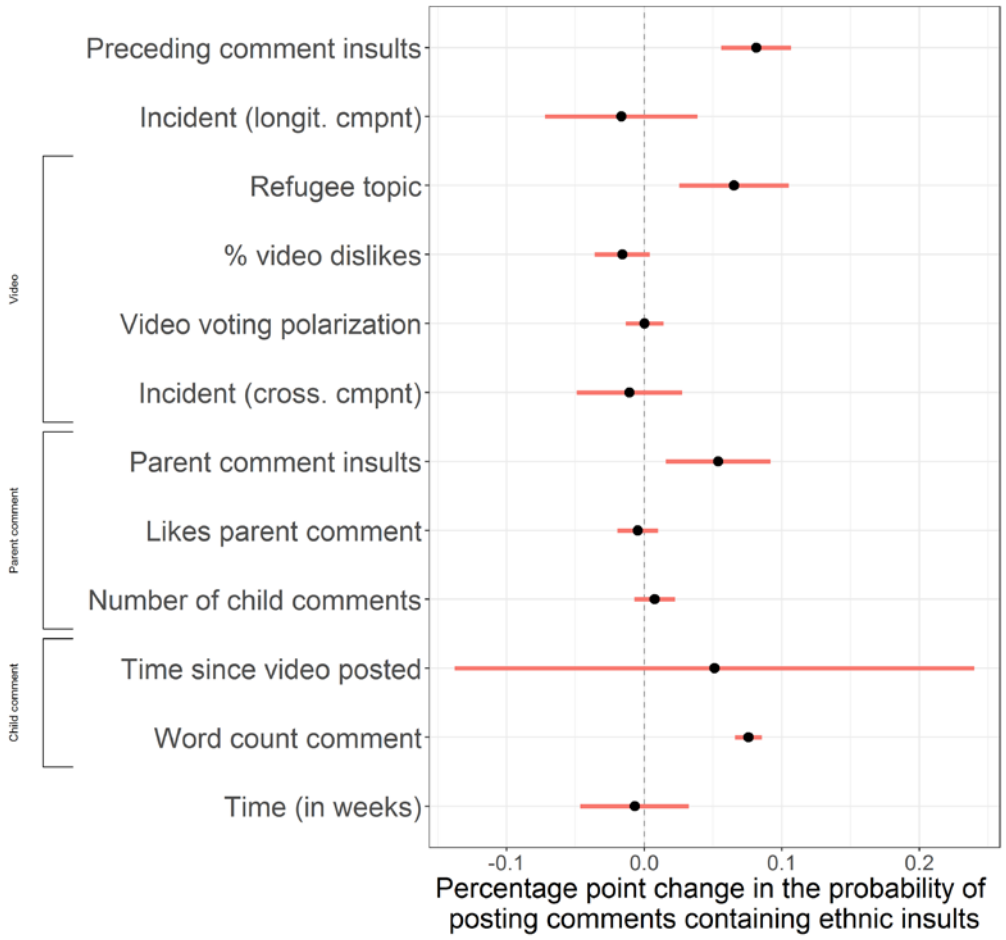
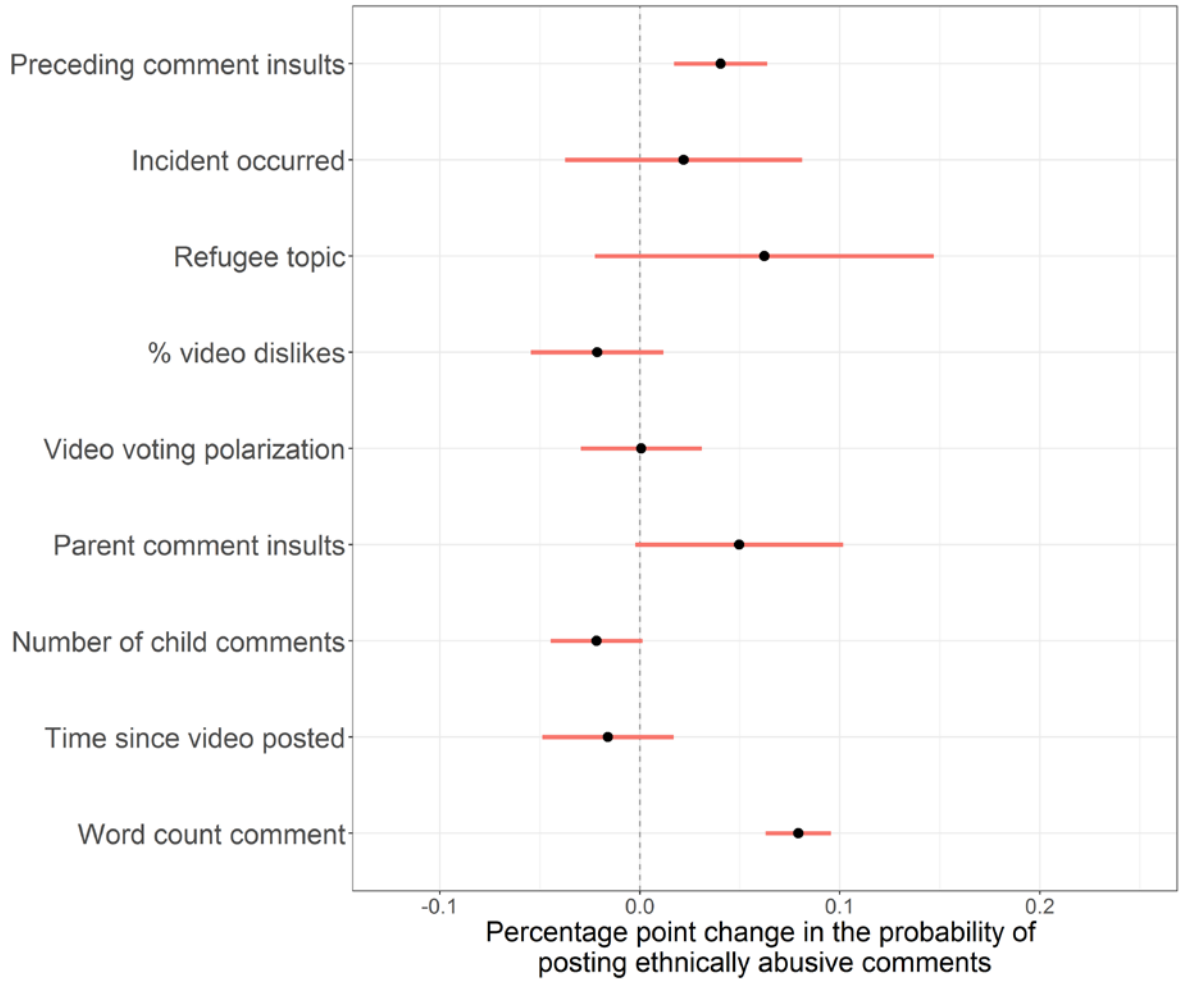
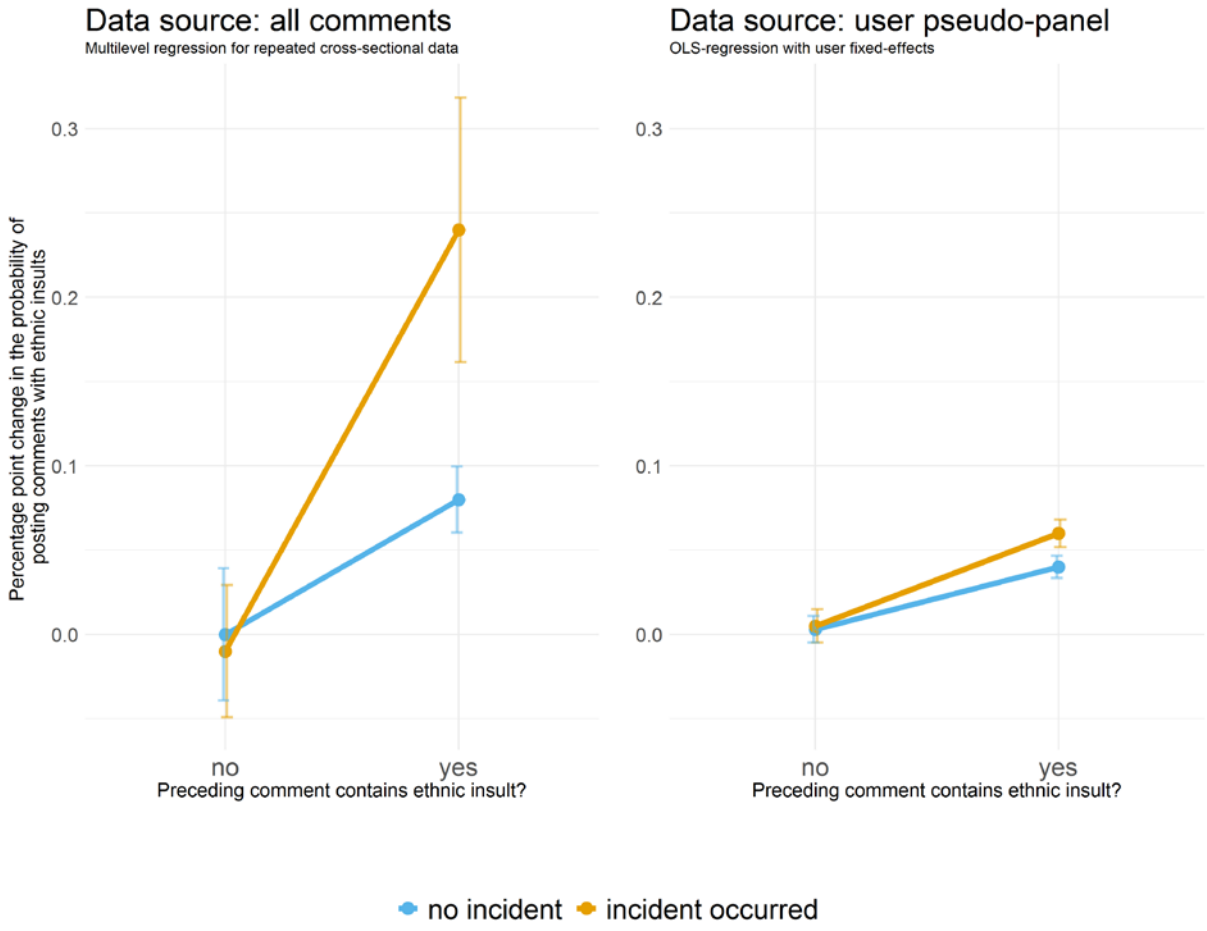


Figure 4: Ethnic insults in YouTube videos of German political talk shows (June 2015-April 2017), user fixed effects OLS regression (nUsers=251, nComments=1,275).



Note: Adjusted R^2=0.17

Figure 5: The contagious effect of preceding comments containing ethnic insults is dependent on the temporal context. Separately for the full sample and the user pseudo-panel.



Appendix

Tables

Table A1: Custom dictionaries

	General insults	Ethnic markers
English	Ass, scum, stupid, bitch, brood, dumb, bottom, fuck, bastard, bigot, retarded, dazzler, clown, dirt, idiot, dilettante, cunt, freak, fascist, grimace, gay, liar, whore, ugly, hater, ignorant, puke, sick, shit, lesbian, loser, licker, mattress, monster, trash, mongo, garbage, nazi, racist, piss, homeless, parasite, pedo, psycho, shit, pig, shabby, faggot, troll, crazy, jerk, traitor	Arab, asylum, refugee, börek, shawarma, islam, muslim, jew, koran, camel, head scarf, kurdish, mohamed, nigger, pole, osman, rapefugee, turk, carpet, terrorist, goat, gypsy.
German	Arsch, abschaum, asshole, bodensatz, blöd, bums, bitch, brut, blende, bescheu, bastar, bigot, beschränk, clown, däm, dumm, dreck, depp, debil, dilet, fick, fotz, fuck, fratz, fresse, freak, fascho, faschist, hartzer, häuch, homo, hure, halunk, häßl, hetzer, gutmensch, gauner, gestört, gesind, idiot, ignoran, jammerla, kotz, kakerlak, krank, kack, lachnummer, lesbe, lügn, loser, luder, lecker, matratz, monster, müll, minderwertig, mongo, nutte, nazi, rassist, piss, penner, parasite, pöbel, pedo, psycho, prolet, siff, spack, socks, schiss, scheiss, schmarotz, schlampe, schwein, spinner, schäbig, schwucht, troll, trottel, tratsch, tussi, verrück, versager, verrät, wix, wichs, wtf, weichei, zecke	Arab, asyl, börek, döner, flüchtl, goldstück, islami, jude, jüdi, koran, kanak, kopfabschn, kümmel, kamel, kopftuch, kurdi, muslim, moslem, musel, mohamed, neger, nigger, pole, polake, osman, rapefug, russe, türke, teppich, terrorist, volkssch, untermen, ziegen, zigeun

Note: German words are presented in their stemmed form, that is, the way they were used to identify swear words. Some German words are loosely translated into English, some words (e.g. stupid) have numerous variations not listed here (e.g., dumm, dämlich, blöd, etc.) and some words are virtually meaningless in English without their context (i.e., “Gutmensch”: good person, used mostly by right-wing proponents to denigrate those who chose to help refugees or are generally supportive of refugees for humanitarian reasons) and are therefore not listed.

Table A2: Descriptive statistics

	Range	**Mean**	**sd**	**Level**
Ethnic insults	0-1	0.08		Comment
Incident (longitudinal)	-1 - +1	0.00	0.31	Video-time
Incident (cross-sectional)	0-1	0.27	0.32	Video
Preceding comment insults (ethnic)	0-1	0.19		Comment
Refugee topic	0-1	0.85		Video
Parent comment insults	0-1	0.10		Parent
Number of Likes parent comment	0-64	10.95	12.95	Parent
Number of child comments	1-20	9.63	6.79	Parent
Time since video posted (in days)	0.13-727.22	238.73	145.60	Comment
Word count comment	1.00-832.00	48.00	68.40	Comment
% video dislikes	0-1	0.50	0.24	Video
Video voting polarization	0-0.5	0.03	0.04	Video
Time (in weeks)	0-104	34.39	20.87	Comment