Reconsidering Evidence of Moral Contagion in Online Social Networks

Jason W. Burton[1]*, Nicole Cruz[2], and Ulrike Hahn[1]

[1] Department of Psychological Sciences, Birkbeck, University of London, London, WC1E 7HX, UK

[2] School of Psychology, University of New South Wales, Sydney, 2052, AU

* Corresponding author: Jason W. Burton (jburto03@mail.bbk.ac.uk; https://orcid.org/0000-0002-6797-2299)

Abstract

The ubiquity of social media use and the digital data traces it produces has triggered a potential methodological shift in the psychological sciences away from traditional, laboratory-based experimentation. The hope is that by using computational social science methods to analyze large-scale observational data from social media, human behavior can be studied with greater statistical power and ecological validity. However, current standards of null hypothesis significance testing and correlational statistics seem ill-suited to markedly noisy, high-dimensional social media datasets. We explore this point by probing the moral contagion phenomenon, whereby the use of moral-emotional language increases the probability of message spread. Through out-of-sample prediction, model comparisons, and specification curve analyses, we find that the moral contagion model performs no better than an implausible XYZ contagion model. This highlights the risks of using purely correlational evidence from large observational datasets and sounds a cautionary note for psychology's merge with big data.

Introduction

The digitalization of society raises many substantive questions (e.g., [1–3]). At the same time, however, it provides unmistakable methodological opportunities for social science research. For all of the interactions that take place online, such as communications between social media users, digital data traces are left behind. Not only do these data traces capture naturalistic behaviors, but the sheer scale and variety of such data means theories of, for example, collective reasoning and opinion dynamics, can now be readily observed quantitatively "in the wild"[4–6]. With more social science domains recognizing this trend and utilizing large-scale social media data to test theories of human behavior[7], there is a pressing need for researchers to better understand the strength of inferences made from such data, and for methodological standards to be queried.

One high-profile study that combines social media data and psychological theory recently presented findings of a "moral contagion" effect[8]. In the study, Brady et al. (2017) apply a dictionary-based text analysis procedure to quantify moral-emotional language in hundreds of thousands of tweets capturing the naturally-occurring communications of Twitter users. By then fitting a regression model and performing a series of robustness checks, they show that the mere presence of moral-emotional words increases messages' retweet counts by a factor of 20%, regardless of the messages' informational quality[8]. The implications of this moral contagion phenomenon, where the exposure to moral emotions shapes the diffusion of information, are undoubtedly significant. Invoking

morality in reasoning has previously been shown to harden existing belief structures, delegitimize authority, and, in extreme cases, dehumanize opposing perspectives[9,10]. While injections of moral reasoning into discourse can be beneficial — providing shared identities and guiding ethical behavior — the introduction of unnecessary moralization and its emotional underpinnings may jeopardize rational debate. It is for this reason that moral justifications carry weight in some domains but not others. For example, loading an argument with moral-emotional language might be an effective strategy in a debate over social policy and human rights, yet that same strategy is likely to be penalized in an argument over mathematics. However, if moral contagion is as widespread and domain-general as Brady et al. (2017) suggest, then it seems plausible that sentiments about where moralization is appropriate are changing. This also suggests that we are susceptible to new forms of political persuasion online. As Brady et al. (2017) conclude, "it seems likely that politicians, community leaders, and organizers of social movements express moral emotions...in an effort to increase message exposure and to influence perceived norms within social networks" (p. 7316). Beyond this substantive contribution, the authors also recognize the methodological implications of their study, because "in comparison with laboratory-based studies, the social network approach offers much greater ecological validity" (8, p. 7317).

Brady et al. (2017) is one example of what is an ongoing methodological shift in the psychological sciences (also see, e.g., [11–13]) , whereby statistical analyses of large-scale digital data traces — namely, social media data — are used as the basis for inferences about human emotions, behaviors, and motivations. But digital data traces produced by social media users are inherently noisy and high-dimensional. In contrast to the "custom-made" data generated via controlled experimentation, material harvested from online platforms is usually not created with research in mind[14]. Social media data can be ambiguous, confounded by proprietary algorithms and restricted access, and unrepresentative of wider populations, which may limit the generalizability of findings between platforms and between online and offline populations[14–16]. These documented observations may be less problematic if one's research objective concerns itself only with behavior on a given platform itself; however, in the absence of agreed upon methodological standards for handling social media data, the space for "researcher degrees of freedom"[17] is particularly vast. This means that conclusions from analyses of observational social media data alone may face deeper issues, insofar as they are intended to teach us something about real human behavior.

In this article, we probe the finding of moral contagion, illustrating possible methodological pitfalls that might be encountered when standard practices of null hypothesis significance testing are applied to large-scale social media datasets. How robust is correlational evidence from large-scale

observational data? What inferences and generalizations can be made from such evidence? Answering these questions seems crucial for psychology's merge with "big data."

Results

Out-of-Sample Prediction

The diffusion of information in social networks has been likened to a biological pathogen, spreading from person to person through direct contact. For a behavior, psychological state, or other condition to qualify as a simple social contagion, the probability of the condition being adopted by an individual should increase monotonically with the number of times that individual is exposed to said condition[18]. In the case of moral contagion, moral-emotional words (e.g., kill, protest, compassion) are considered to be the "contagious" cue because their presence is presumed to be a central factor in an individual's decision to retweet (or diffuse) the message in which it is included. Based on this logic, moral contagion should be present in other corpora of tweets pertaining to contentious, politicized topics. To test this proposal, we recreated Brady et al.'s (2017) methodology and applied it to other Twitter corpora spanning a variety of socio-political issues and events.

Using the dictionary-based text analysis of Brady et al. (2017) to quantify distinctly moral, distinctly emotional, and moral-emotional language (see Supplementary Information, Section 1.2 for details), we tested the influence of language use on message diffusion across six corpora of tweets that capture the naturally-occurring communications among users (four of these corpora were pre-existing). Each corpus pertained to a specific issue or event: the COVID-19 pandemic (n = 172,697), the #MeToo movement[19] (n = 151,035), the #MuellerReport investigation (n = 39,068), the #WomensMarch protest[20] (n = 3,778), the announcement of the 2016 EU Referendum result in the United Kingdom (Brexit)[21] (n = 5,660), and the 2016 US Presidential Election (n = 8,233; this corpus contained only "viral" tweets that received more than 1,000 retweets)[22] (see Supplementary Information, Section 1.1 for details on each corpus). Diffusion was measured as the sum of a message's retweet count as captured in the metadata and the number of times that message's text appeared in a corpus. Identical messages were then collapsed into a single observation with other relevant metadata from the earliest posting (e.g., the number of followers a message poster has; whether the post included URLs, an image, or video media). This approach avoids penalizing retweet chains, which are important indicators of diffusion on Twitter, while also accounting for unconventional retweets where a user copies and pastes someone's message rather than clicking the retweet button. With diffusion as our dependent variable and the three language measures as

predictors, we then followed Brady et al. (2017) in fitting a negative binomial regression model to each dataset (henceforth referred to as the "main moral contagion model"). The presence of contagion was determined by exponentiating the regression coefficients of each predictor (i.e., distinctly emotional, distinctly moral, and moral-emotional language) to generate incidence rate ratios (IRR) — the most central measure being moral-emotional language's IRR. Note that as a ratio measure, IRRs greater than 1.00 signify a positive contagion effect (e.g., IRR = 1.10 suggests a 10% increase in diffusion), and vice versa.

Prior to analyzing our corpora, we checked our model specifications by reanalyzing Brady et al.'s (2017) cleaned data, which they have made available online. Across the three corpora comprising 313,002 analyzable tweets spanning three topics (same-sex marriage, n = 29,060; gun control, n = 48,394; climate change, n = 235,548), our analysis reproduced their findings. Moral-emotional language was significantly associated with an increase in retweets in each corpus when covariates were controlled for (same-sex marriage, IRR = 1.17, $p < 0.001$, 95% CI = 1.09, 1.27; gun control, IRR = 1.19, $p < 0.001$, 95% CI = 1.14, 1.23; climate change, IRR = 1.24, $p < 0.001$, 95% CI = 1.22, 1.27), and in two out of three corpora when covariates were not controlled for (same-sex marriage, IRR = 1.08, $p = 0.059$, 95% CI = 0.99, 1.18; gun control, IRR = 1.36, $p < 0.001$, 95% CI = 1.30, 1.42; climate change, IRR = 1.15, $p < 0.001$, 95% CI = 1.12, 1.17). However, these results did not consistently generalize across the six corpora we analyzed.

Taking Brady et al.'s main moral contagion model, as well as the nested single-variable model in which only moral-emotional language is used as a predictor, we found moral contagion to be present in only two of six corpora before controlling for covariates: COVID-19 tweets (IRR = 1.15, $p < 0.001$, 95% CI = 1.11, 1.18) and #MuellerReport tweets (IRR = 1.28,$p < 0.001$, 95% CI = 1.16, 1.42). In the four pre-existing corpora, moral-emotional language either had no significant relationship with message diffusion or had a negative effect where moral-emotional language predicted a decrease in diffusion (Table 1). While we could not control for the same covariates as Brady et al. (2017) and were therefore unable to provide direct replications in the four pre-existing corpora due to missing metadata, we did so in the COVID-19 and #MuellerReport corpora (we do this to aid comparison with Brady et al.'s (2017) original results; however, we strongly caution against basing one's interpretation of these results on covariates – see section on "covariates, outliers, and the analytical multiverse"). Once Brady et al.'s (2017) chosen covariates were controlled for in the regression model to provide a direct replication of the original analysis, the significant association between moral-emotional words and message diffusion remained in the #MuellerReport tweets (IRR = 1.27, *p*

< 0.001, 95% CI = 1.16, 1.40), but no statistically significant relationship was observed in the COVID-19 tweets (IRR = 1.01, p = 0.320, 95% CI = 0.99, 1.04).

The limits of correlational data

The inconsistent results of out-of-sample prediction tests using similar procedures as Brady et al. (2017) point toward the limitations of purely correlational data. The inherent difficulty of distinguishing true causal contagion from confounding network homophily has been noted in detail elsewhere (e.g., [23,24]). But large sets of observational data carry even more fundamental risks of spurious correlation and endogeneity. To demonstrate this, we conducted a follow-up analysis in the spirit of Hilbig (2010) who cautioned against correlational data as a sole source of evidence for heuristic use in judgement and decision-making tasks[25]. Specifically, we created an absurd factor for illustrative purposes, what we call XYZ contagion, and tested whether the number of X's, Y's, and Z's included in messages' text predicted diffusion (note that we were unable to test for XYZ contagion in Brady et al.'s (2017) original data because their raw data did not include metadata retweet counts, which meant that our analysis scripts could not be properly applied).

Our analysis found XYZ contagion to be present in four of our six corpora such that the presence of the letters X, Y, and Z predicted an increase in message diffusion: COVID-19 tweets (IRR = 1.08, p < 0.001, 95% CI = 1.07, 1.08), #MeToo tweets (IRR = 1.13, p < 0.001, 95% CI = 1.12, 1.15), #MuellerReport tweets (IRR = 1.12, p < 0.001, 95% CI = 1.10, 1.14), and the 2016 US Election tweets (IRR = 1.01, p = 0.030, 95% CI = 1.00, 1.03). While there was no positive relationship between the presence of X, Y, and Z and message diffusion in the #WomensMarch and Post-Brexit tweets, the finding that XYZ contagion passes a key test of robustness, out-of-sample prediction, demonstrates the potential of large-scale social media datasets to contain spurious correlations (Table 1; also see Supplementary Information, Section 2.3 for a bootstrap resampling analysis).

In addition, we calculated Akaike Information Criteria (AIC) as measures of model adequacy and found that our model of XYZ contagion actually outperforms the main, multi-variable moral contagion model in two of the six corpora (Table 1). We further tested the XYZ contagion model against the single variable moral contagion model such that the predictive value of the count of letters X, Y, and Z was compared to the count of moral-emotional words in isolation. This analysis revealed that the count of letters X, Y, and Z was in fact a better predictor of message diffusion than moral-emotional words in five out of six corpora, despite being nonsensical (Table 1).

Covariates, outliers, and the analytical multiverse

Out-of-sample prediction tests and model comparisons demonstrate how social media datasets may be susceptible to unfounded correlations. However, we need to consider the influence of outliers and covariates in more detail, which are indeed sensible and widely-recognized checks that can and have been put in place to guard against spurious results. But as we show next, in the context of social media data, neither of these are sufficient to solve the problems identified here, facing both methodological and conceptual limitations.

Regarding outliers, the problem is that social media data are a typical case of fat-tailed distribution, and it is unclear how "outlier" should be defined. The prevalence of extreme values (e.g., a tweet garnering 100,000 retweets when the median is 0) is likely a constitutive feature of the dataset, rather than a bug or error to be neglected. Consequently, decisions on outliers are seemingly arbitrary. For example, consider a traditional psychology experiment measuring reaction times in the lab. Outliers in this case are readily identifiable: A reaction time that is ten times the mean indicates that a participant was not paying attention, had not read the instructions, or the data was entered incorrectly. Yet, in the domain of social media, there is no such judgement that can be made. That a message may be retweeted zero, one, or 100,000 times is in fact an intrinsic part of the paradigm. What does it mean if, in a study of message sharing, the top ten or one hundred most shared messages determine what statistical results are retrieved from a corpus of hundreds of thousands of messages? Are these observations to be excluded, or are they indicative of a recipe for going viral?

Covariates might be considered even more important. Indeed, there is a wide range of potential covariates that plague social media data, relating to both the content of messages and the accounts of message posters. Specifically relating to Twitter, it has previously been shown that the presence of hashtags and URLs in a message, the number of followers and followees a message poster has, and the age of the message poster's account all influence retweet rates[26]. There are also questions around the potential need to account for the influence of automated and semi-automated bots[16,27,28]. Despite existing literature highlighting these covariates, the controls that researchers put in place are often inconsistent, even when the hypotheses in question are relatively similar. For example, consider three studies investigating the role of emotion in message sharing on Twitter: Stieglitz and Dang-Xuan (2013) control for the number of hashtags a tweet contains, the presence of URLs, the number of followers a message poster has, and the number of tweets a user has posted during the sampling period[29]; Ferrara and Yang (2015) excluded tweets containing URLs or media (i.e., a photo or video)[30]; and Brady et al. (2017) control for the number of followers the message poster has, whether media or URLs are present in a tweet, and whether the message poster is "verified" (a status indicating that the user is a celebrity or public figure). Not only do these studies

identify different covariates, but they also control for them in different ways. For instance, where Ferrara and Yang (2015) excluded tweets containing URLs and media, Brady et al. (2017) input these covariates as binary variables in a regression. While each study's controls are certainly defensible, this points to another problem: any given set of controls will not be exhaustive and there is no agreed upon standard for what controls must be made to separate a publishable finding from a coincidental statistic; and even more fundamentally, against what ground truth could these methodological practices be evaluated?

Taken together, the ambiguity surrounding outliers and covariates highlights the increased "researcher degrees of freedom"[17] in analyses of social media data. That is, researchers must make many arbitrary analytical decisions when collecting, processing, and analyzing the data. While this is not unique to social media data or any type of digital data traces, it may be especially consequential in this context. To investigate how decisions on covariates and outliers influence the moral contagion and XYZ contagion results, we conducted specification curve analyses (SCA)[31] on our three largest corpora (COVID-19, #MeToo, and #MuellerReport). In short, SCA is a way to make analytic flexibility transparent by running all justifiable model specifications (e.g., what covariates to control for, what data subsets to analyze, what independent variable to assess, etc.), and then making joint inferences across the results of all these specifications[31]. SCA is closely related to the concept of a "garden of forking paths"[32] and "multiverse analysis"[33], and serves to clarify the fragility or robustness of statistical findings by identifying which analytical choices they hinge on.

For our SCA, we consider the results of negative binomial regression specifications with either the number X's, Y's, and Z's or the number of moral-emotional words in a tweet predicting diffusion, with or without controlling for covariates, and with or without the removal of (arbitrary) increments of outliers (the tweets with the top 10, 100, and 1,000 diffusion counts). The covariates we consider are the number of distinctly moral words, the number of distinctly emotional words, and the number of characters in a tweet, the number of followers a message poster has, whether the message poster's account is verified, and whether media, URLs, and hashtags are present (binary). Because the #MeToo corpus is a preexisting dataset that was not collected by the authors of the present study, not all of the relevant metadata is included and only some of the covariates could be considered. Figure 1 displays the outcome (unstandardized regression coefficient) of each model specification (x-axis) when fitted to each corpus as three, vertically-aligned points corresponding to the independent variable, covariates, and outliers accounted for (y-axis). We then plot these outcomes as specification curves in Figure 2, visualizing how negative, positive, and nonsignificant moral contagion effects can be retrieved, depending on the chosen corpus and model specification

(also see Supplementary Figures 2-5 for SCA applied to Brady et al.'s original corpora). The specification curves also allow for comparative evaluations between moral contagion and XYZ contagion. Namely, we observe that while the median regression coefficient across model specifications with moral-emotional words as the independent variable is positive in the COVID-19 (n = 40, median B = 0.18, SD = 0.08) and #MuellerReport corpora (n = 39, median B = 0.10, SD = 0.13), it is negative in the #MeToo corpus (n = 28, median B = -0.02, SD = 0.08). Meanwhile, the median regression coefficient across model specifications with the number of X's, Y's, and Z's as the independent variable is positive in all three corpora (COVID-19, n = 39, median B = 0.07, SD = 0.05; #MeToo, n = 28, median B = 0.04, SD = 0.06; #MuellerReport, n = 39, median B = 0.05, SD = 0.05). This could be taken to suggest that the XYZ contagion effect is, if anything, more stable than the moral contagion effect across theoretically-justifiable model specifications in the three corpora addressed here. Of course, we strongly doubt that the letters X, Y, and Z play a central role in shaping the diffusion of information on Twitter. What our analyses show, however, is that the evidence of moral contagion provided by Brady et al. (2017) seems to be virtually indistinguishable from our atheoretical XYZ contagion effect, regardless of whether it is framed as a causal or correlational effect.

Discussion

Out-of-sample prediction, model comparisons, and SCA question the evidence that a meaningful moral contagion effect has been identified on Twitter. To be clear, moral contagion may very well exist, as lab-based work seems to support[34], but our results caution against basing such a conclusion on large-scale, observational data alone. They also caution against the idea that such data provide stronger evidence than lab-based studies due to greater ecological validity. Not only does our analysis challenge the moral contagion hypothesis, but, perhaps most worryingly, it shows that current methodological standards can support patently absurd models, such as the XYZ contagion. One limitation of our analysis is that it is indeed possible to hypothesize why the XYZ contagion might exist after seeing our results (e.g., perhaps X's, Y's, and Z's are attention-grabbing because they are infrequently used). However, there is no reason to believe that the presence of these letters is causally relevant *a priori* and there is currently no evidence to suggest such a theory. While one might expect such causally irrelevant factors to be randomly distributed, there is no guarantee that they do not exhibit some artefactual, spurious correlation with the target phenomenon of interest. Yet crucially, the analyst has no way of telling in advance what state of affairs they will face. For analyses of digital data traces collected from social media platforms to effectively inform psychological inquiry, we make two suggestions for future research utilizing such data: (1) do not

settle for correlational evidence alone, and (2) make the consequences of analytic flexibility transparent.

Both the fragility of moral contagion and the seeming "success" of XYZ contagion in our data highlight how the conclusions afforded by standard statistical procedures, like linear regression models and significance testing, are limited when applied to large-scale social media datasets. While correlational evidence can be informative (e.g., for predictive purposes), this overlooks the crucial point of why findings such as the moral contagion phenomenon are typically interesting. Arguably, the correlational findings of moral contagion are interesting precisely where they seem to be indicative of a meaningful causal relationship[35]. This is why it would be highly unlikely that any academic journal would publish a paper on XYZ contagion. It thus seems necessary for researchers interested in understanding human behavior to either triangulate correlational findings with data from controlled experimentation (e.g., [36,37]); apply alternative statistical techniques, such as structural equation modelling (SEM)[38] or directed acyclic graphs (DAGs)[35]; or use other design methods for causal inference with observational data, if large-scale observational data is to be relied upon.

Our analysis also highlights the need to address analytic flexibility when utilizing social media data. The SCA results presented show how justifiable decisions on covariates and outliers are empirically consequential, capable of giving rise to directly conflicting results on the same predictive relationship in the same dataset. Yet our demonstration only scratches the surface of the analytical "multiverse" that researchers must navigate when handling social media data. For instance, text-as-data research such as that examined in the present work requires heavy data preprocessing, for which there is no agreed upon standard. In analyzing tweets, one may or may not decide to employ stemming, lemmatization, remove "stop words", remove usernames and hashtags, disambiguate homographs with part-of-speech tagging (e.g., "be *kind* to your dog" vs. "what *kind* of dog is that?"), and so on. While seemingly mundane, these preprocessing decisions can lead substantively different interpretations of the data to emerge[39]. The same can also be said of feature engineering. For example, the decision to use a dictionary (or bag-of-words) approach versus a machine learning strategy for text classification can lead to different measurements of moral expressions within the same corpus, and classification performance can vary across contexts[40]. While a logistic regression model fitted with Brady et al.'s (2017) dictionaries seems to be a good predictor of human judgements of moral expression in tweets related to #MeToo (AUC = 83.2%), it is essentially as good as random when applied to a corpus containing hate speech messages (AUC = 51.7%) (see Supplementary Information, Section 2.1 for more analysis of the Moral Foundations Twitter

Corpus[40]). At present, the focal strategy for managing analytic flexibility is pre-registration, but this seems ineffective for the issues raised here. Pre-registering an analysis plan might ensure researchers commit to a chosen analytical pathway and guard against "p-hacking," but given the underlying multiverse of divergent but theoretically-defensible results, this is not enough to guarantee that the specific results retrieved are ultimately informative. While there is indeed a longstanding tradition in the social sciences to consider alternative model specifications as a check of robustness, methods like SCA[31] should be encouraged so as to make this tradition more transparent and exhaustive, and to better display exactly which analytical decisions are responsible for potentially conflicting results.

While the use of observational "big data" is relatively new to the psychological sciences, the obstacles outlined here are not particularly novel in other fields. For instance, large longitudinal datasets have been integral to the study of public health and epidemiology, where it has previously been shown that the standard use of regression models can produce implausible findings, such as statistics that suggest acne, headaches, and height are "contagious"[41]. If analyses of social media and other digital data traces are to contribute to the study of human psychology, it seems unlikely that the standard practices of null hypothesis significance testing and robustness checks will suffice. As demonstrated here, the inferences and generalizations that can be made from purely correlational findings in observational social media data can sometimes be remarkably fragile.

Methods

A total of 380,471 unique tweets were analyzed in the present work (COVID-19, n = 172,697; #MeToo, n = 151,035; #MuellerReport, n = 39,068; 2016 US Election, n = 8,233; Post-Brexit, n = 5,660; #WomensMarch, n = 3,778), not including the re-analysis of Brady et al.'s (2017) corpora comprising 313,002 unique tweets. No statistical methods were used to pre-determine sample sizes, but we sought to compile a collection of corpora that is comparable in size to that used in Brady et al. (2017). Four of the corpora were retrieved from open data repositories (#MeToo, #WomensMarch, Post-Brexit, and 2016 US Election) and two were collected by the authors from the Twitter REST API with the "rtweet" package in R (COVID-19 and #MuellerReport). For a detailed explanation of the collection parameters, justification for inclusion, and references see Supplementary Information, Section 1.1. The data pre-processing procedure was identical in both the analysis of moral contagion and XYZ contagion, which included the removal of URLs and username tags from each message's text.

All statistical tests reported are two-tailed. To assess for the presence of moral contagion and XYZ contagion, we followed Brady et a. (2017) and modelled the data with negative binomial regressions with maximum likelihood estimation to best handle the overdispersed count data being analyzed[42]. To compare the quality of each candidate model, *i*, we used the AIC measure and calculated the difference between each model, Δ_i (AIC), as well as the corresponding Akaike weights[43], w_i (AIC) (Table 1). We then investigated how analytical flexibility might allow for conflicting results to arise with specification curve analyses (SCA)[31]. To do so, we used the "specr"[44] package in R.

Data availability

The data files analysed in this study are publicly available. For the COVID-19 and #MuellerReport corpora, which were specifically collected for this study, the tweet IDs have been made available on OSF but not the tweet text due to restrictions set by Twitter. To access the pre-existing corpora, please see references 19-22 in the reference list, or follow the links from the OSF project page directing to the original hosting sites. [https://osf.io/4zjk6/?view_only=a9f977a26f754b2da4c59b2fba29cd50].

Code availability

All scripts used for the analyses presented in this article are available on the OSF project page. [https://osf.io/4zjk6/?view_only=a9f977a26f754b2da4c59b2fba29cd50].

References

1. Tufekci, Z. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. (Yale University Press, 2017).
2. Sunstein, C. R. *#Republic: Divided Democracy in the Age of Social Media*. (Princeton University Press, 2018).
3. Moore, M. *Democracy Hacked: Political Turmoil and Information Warfare in the Digital Age*. (Oneworld Publications, 2019).
4. Lazer, D. *et al.* Computational Social Science. *Science* **323**, 721–723 (2009).
5. Giles, J. Making the links. *Nature* **488**, 448–450 (2012).
6. Conte, R. *et al.* Manifesto of computational social science. *Eur. Phys. J. Spec. Top.* **214**, 325–346 (2012).

7. Edelmann, A., Wolff, T., Montagne, D. & Bail, C. A. Computational Social Science and Sociology. *Annu. Rev. Sociol.* **46**, 61-81 (2020).
8. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci.* **114**, 7313–7318 (2017).
9. Crockett, M. J. Moral outrage in the digital age. *Nat. Hum. Behav.* **1**, 769–771 (2017).
10. Ben-Nun Bloom, P. & Levitan, L. C. We're Closer than I Thought: Social Network Heterogeneity, Morality, and Political Persuasion. *Polit. Psychol.* **32**, 643–665 (2011).
11. De Choudhury, M., Counts, S. & Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* 3267 (2013).
12. Tumasjan, A., Sprenger, T. O., Sander, P. G. & Welpe, I. M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* 178–185 (2010).
13. Garcia, D. & Rimé, B. Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack. *Psychol. Sci.* **30**, 617–628 (2019).
14. Salganik, M. *Bit by Bit: Social Research in the Digital Age*. (Princeton University Press, 2017).
15. Tufekci, Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* 505–514 (2014).
16. Ruths, D. & Pfeffer, J. Social media for large studies of behavior. *Science* **346**, 1063–1064 (2014).
17. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
18. Hodas, N. O. & Lerman, K. The Simple Rules of Social Contagion. *Sci. Rep.* **4**, 4343 (2014).
19. Turner, A. 390,000 #MeToo Tweets. *data.world* (2018).
20. Adhokshaja, P. #Inauguration and #WomensMarch. *data.world* (2017).
21. Parker, C. Brexit Tweets from the morning of its announcement. *Mendeley Data* (2017).
22. Amador, J., Oehmichen, A. & Molina-Solana, M. Fakenews on 2016 US elections viral tweets (November 2016 - March 2017). *Zenodo* (2017).
23. Shalizi, C. R. & Thomas, A. C. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociol. Methods Res.* **40**, 211–239 (2011).
24. Aral, S., Muchnik, L. & Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci.* **106**, 21544–21549 (2009).

25. Hilbig, B. E. Reconsidering “evidence” for fast-and-frugal heuristics. *Psychon. Bull. Rev.* **17**, 923–930 (2010).
26. Suh, B., Hong, L., Pirolli, P. & Chi, E. H. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. in *2010 IEEE Second International Conference on Social Computing* 177–184 (2010).
27. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
28. Kollanyi, B., Howard, P. N. & Woolley, S. C. *Bots and automation over Twitter during the U.S. election*. (2016).
29. Stieglitz, S. & Dang-Xuan, L. Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *J. Manag. Inf. Syst.* **29**, 217–248 (2013).
30. Ferrara, E. & Yang, Z. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Comput. Sci.* **1**, e26 (2015).
31. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208-1214 (2020).
32. Gelman, A. & Loken, E. The Statistical Crisis in Science. *Am. Sci.* **102**, 460–466 (2014).
33. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
34. Brady, W. J., Gantman, A. P. & Van Bavel, J. J. Attentional capture helps explain why moral and emotional content go viral. *J. Exp. Psychol. Gen.* **149**, 746–756 (2020).
35. Rohrer, J. M. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Adv. Methods Pract. Psychol. Sci.* **1**, 27–42 (2018).
36. Mooijman, M., Hoover, J., Lin, Y., Ji, H. & Dehghani, M. Moralization in social networks and the emergence of violence during protests. *Nat. Hum. Behav.* **2**, 389–396 (2018).
37. Dehghani, M. *et al.* Purity homophily in social networks. *J. Exp. Psychol. Gen.* **145**, 366–375 (2016).
38. Westfall, J. & Yarkoni, T. Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE* **11**, e0152719 (2016).
39. Denny, M. & Spirling, A. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Polit. Anal.* **26**, 168–189 (2018).
40. Hoover, J. *et al.* Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Personal. Sci.* **11**, 1057-1071 (2019).
41. Cohen-Cole, E. & Fletcher, J. M. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ* **337**, a2533–a2533 (2008).
42. Hilbe, J. M. *Negative binomial regression*. (Cambridge University Press, 2011).

43. Wagenmakers, E.-J. & Farrell, S. AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **11**, 192–196 (2004).
44. P. Masur & Sharkow, M. *specr: Statistical functions for conducting specification curve analyses (Version 0.2.1)*. (2019).

Acknowledgements

The authors received no specific funding for this work. We thank the Causal Cognition Lab at University College London for their helpful feedback on this work.

Author contributions

J.W.B. and U.H. designed the research, J.W.B. and N.C. analyzed the data, J.W.B wrote the paper and all authors contributed to revisions.

Competing interests

The authors declare no competing interests.

Figure Legends

Figure 1. Qualitative results of specification curve analyses (SCA). (A) COVID-19 corpus. (B) #MeToo corpus. (C) #MuellerReport corpus. Each possible model specification (x-axis) is represented by three vertically-aligned points corresponding to the outliers removed and covariates and independent variable included in the negative binomial regression equation (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. There are fewer specifications (56) in the #MeToo SCA (B) because metadata on some covariates was absent. Of the 80 possible specifications for the COVID-19 and #MuellerReport data, one specification was excluded from the COVID-19 SCA and two specifications were excluded from the #MuellerReport SCA because these algorithms did not converge.

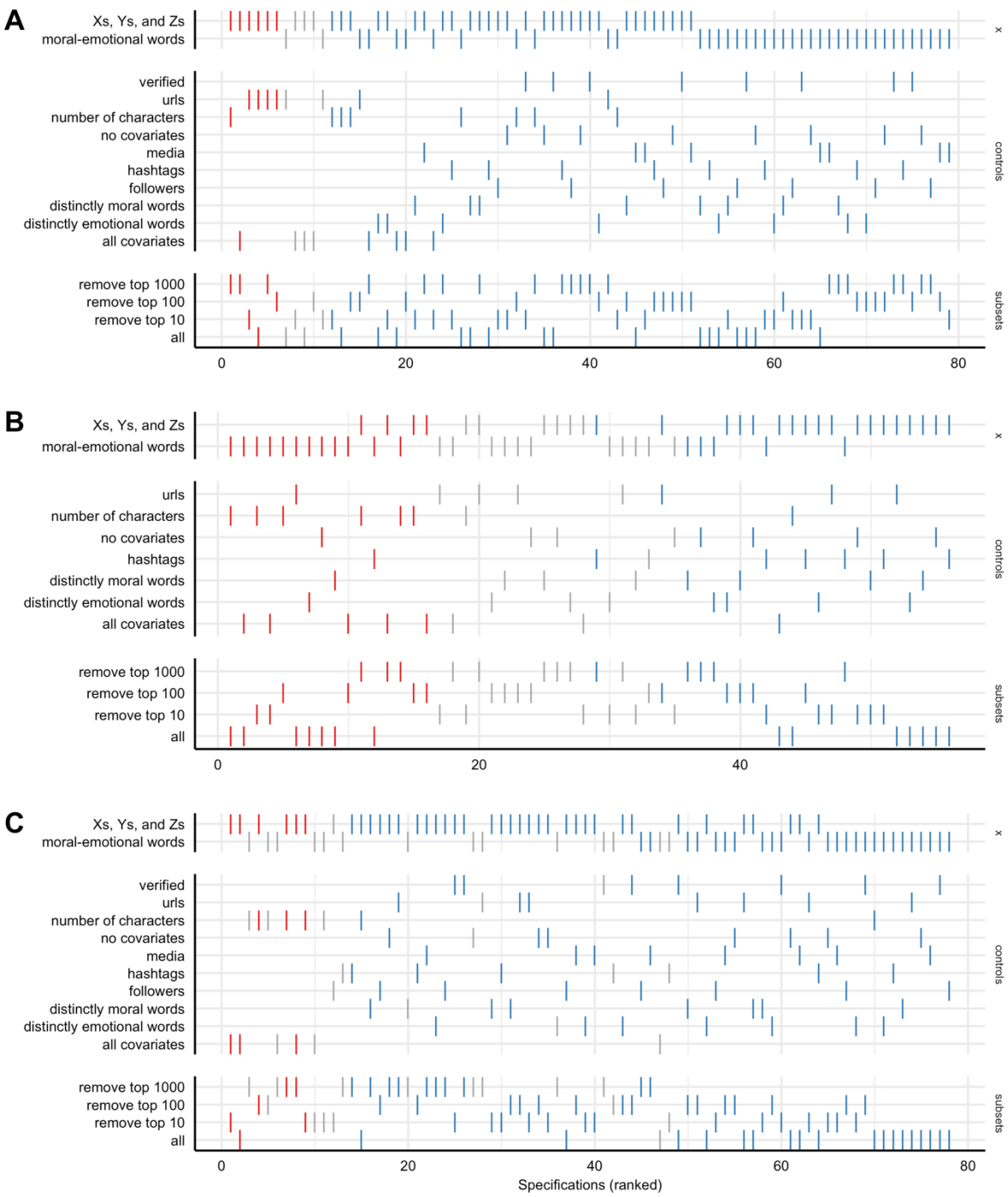
Figure 2. Specification curves for moral contagion (top plots) and XYZ contagion (bottom plots) effects. (A) COVID-19 corpus (moral contagion, n = 40, median B = 0.18, SD = 0.08; XYZ contagion,

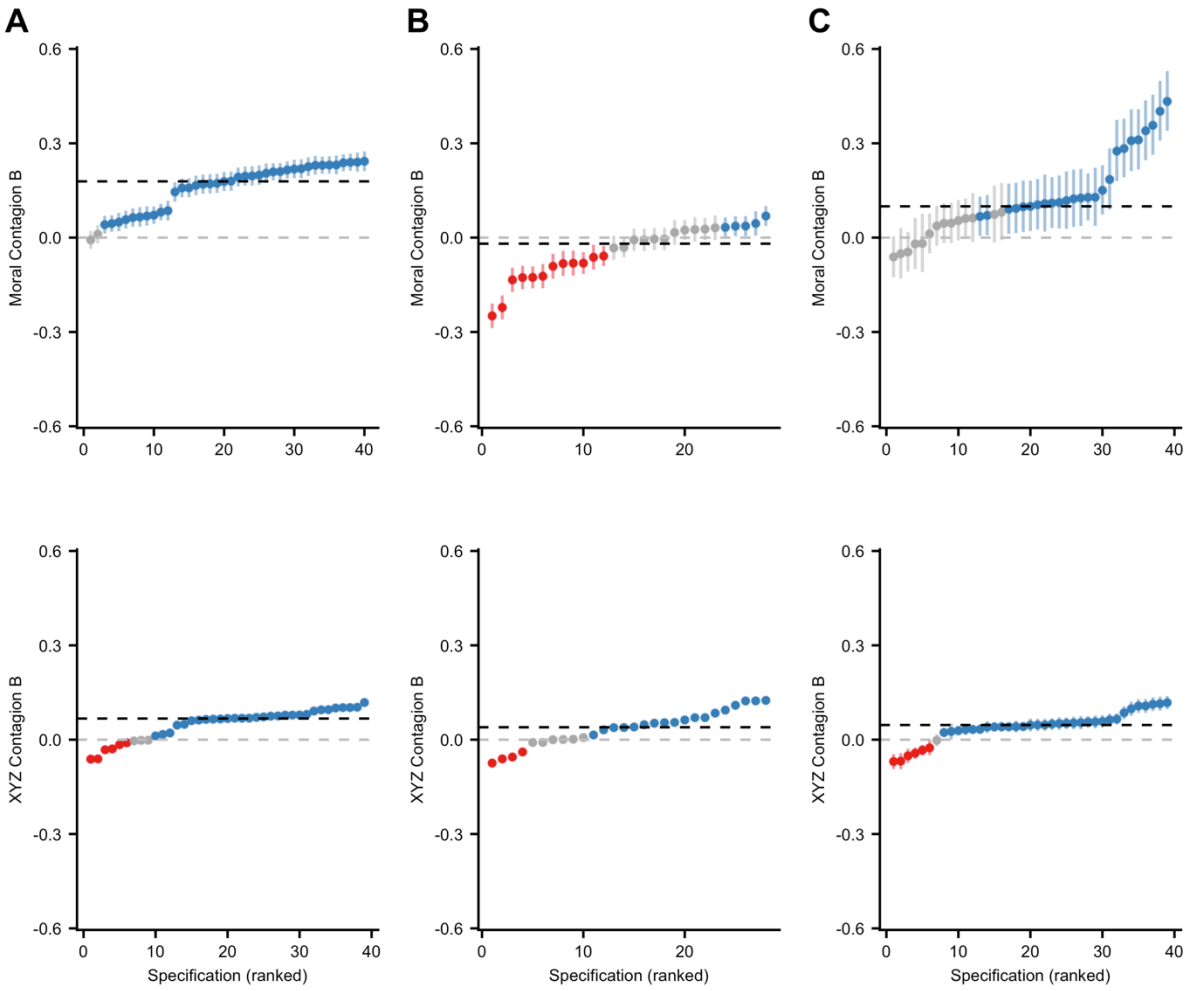
n = 39, median B = 0.07, SD = 0.05). (B) #MeToo corpus (moral contagion, n = 28, median B = -0.02, SD = 0.08; XYZ contagion, n = 28, median B = 0.04, SD = 0.06). (C) #MuellerReport corpus (moral contagion, n = 39, median B = 0.10, SD = 0.13; XYZ contagion, n = 39, median B = 0.05, SD = 0.05). Each model specification (x-axis) is represented by a single point indicating the resulting unstandardized regression coefficient and vertical bars indicating 95% confidence intervals (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. There are fewer specifications in the #MeToo corpus (B) because metadata on some covariates was not recorded. Two specifications in the #MuellerReport corpus (C) and one specification in the COVID-19 corpus (A) are excluded because the algorithm did not converge.

Tables

Table 1. Negative binomial regression model results and comparisons. Incidence rate ratios (IRR) indicate the size of contagion effects in each dataset (for the main moral contagion model only the effect of moral-emotional language is reported), with 95% confidence intervals in brackets and corresponding p-values in the row below. For each model, the differences in AIC with respect to the best candidate is calculated, Δ_i (AIC), meaning that an Δ_i (AIC) equal to zero signals that the corresponding model is the best fit for the given dataset. AIC values are further transformed into Akaike weights, w_i (AIC), which are the conditional probabilities that the model in question, *i*, is the best model given the data and the set of candidate models[43].

	COVID-19	**#MeToo**	**#Mueller Report**	**2016 US Election**	**Post-Brexit**	**#Womens March**
N	172,697	151,035	39,068	8,233	5,660	3,778
			Main Multi-Variable Moral Contagion Model			
IRR	1.15 [1.11, 1.18]	0.91 [0.88, 0.95]	1.28 [1.16, 1.42]	1.02 [0.98, 1.06]	0.89 [0.72, 1.13]	1.01 [0.77, 1.38]
p	< 0.001	< 0.001	< 0.001	0.465	0.370	0.925
Δ_i (AIC)	0.00	138.88	20.66	0.00	0.00	0.00
w_i (AIC)	> .9999	< .0001	< .0001	> .9999	.9995	0.9629
			Single-Variable Moral Contagion Model			
IRR	1.19 [1.15, 1.23]	0.92 [0.89, 0.96]	1.40 [1.28, 1.55]	1.02 [0.98, 1.06]	0.81 [0.66, 1.02]	0.90 [0.68, 1.24]
p	< 0.001	< 0.001	< 0.001	0.337	0.101	0.494
Δ_i (AIC)	690.53	334.28	49.41	36.39	15.39	14.28
w_i (AIC)	< .0001	< .0001	< .0001	< .0001	.0005	0.0009
			XYZ Contagion Model			
IRR	1.08 [1.07, 1.08]	1.13 [1.12, 1.15]	1.12 [1.10, 1.14]	1.01 [1.00, 1.03]	1.00 [0.95, 1.06]	0.89 [0.82, 0.96]
p	< 0.001	< 0.001	< 0.001	0.030	0.998	0.011
Δ_i (AIC)	386.72	0.00	0.00	32.67	18.56	6.87
w_i (AIC)	< .0001	> .9999	> .9999	< .0001	.0001	0.0362


A
Xs, Ys, and Zs
moral-emotional words
x
verified
urls
number of characters
no covariates
media
hashtags
followers
distinctly moral words
distinctly emotional words
all covariates
controls
remove top 1000
remove top 100
remove top 10
all
subsets
0
20
40
60
80
B
Xs, Ys, and Zs
moral-emotional words
x
urls
number of characters
no covariates
hashtags
distinctly moral words
distinctly emotional words
all covariates
controls
remove top 1000
remove top 100
remove top 10
all
subsets
0
20
40
C
Xs, Ys, and Zs
moral-emotional words
x
verified
urls
number of characters
no covariates
media
hashtags
followers
distinctly moral words
distinctly emotional words
all covariates
controls
remove top 1000
remove top 100
remove top 10
all
subsets
0
20
40
60
80
Specifications (ranked)


A
B
C
Moral Contagion B
XYZ Contagion B
0.6
0.3
0.0
-0.3
-0.6
0
10
20
30
40
Specification (ranked)