Emotional Sharing on Social Media:

How Twitter Replies Contribute to Increased Emotional Intensity

Emotional sharing is thought to be a driver of social movements on social media. But how does this work, particularly given how quickly emotions usually decay? In the present research, we test the hypothesis that emotional sharing in twitter replies shapes "appropriate" emotional responses, thereby supporting social movements. In Study 1, we consider a negative context (the Ferguson unrest) and show that replies are more negative than original tweets, and that this process is mainly driven by replies to positive and neutral tweets. We further show that replies that are more negative than original tweets are rewarded by more likes and shares. In Study 2, we consider a positive context (Trump supporters celebrating his election victory) and show that replies are more positive than original tweets, and that this positivity is rewarded by more likes and shares. Combined, these two studies reveal emotional sharing processes that may drive social movements.

Social media are generally thought to have played a key role in the social movements seen in recent social movements. These include the Arab Spring (McGarty, Thomas, Lala, Smith, & Bliuc, 2014; Tufekci & Wilson, 2012), the Occupy movement (Juris, 2012; Smith, Gavin, & Sharp, 2015), and Black Lives Matter (Garza, 2015). What mechanisms underlie the potency of social media in the development of social movements?

One candidate mechanism is emotional sharing. People experience strong emotions in response to important social causes, and these emotions often motivate people to share their emotions (for a review, see Rimé, Finkenauer, Luminet, Zech, & Philippot, 1998). Indeed, a growing literature suggests that emotions expressed on social media often spread quickly (Alvarez, Garcia, Moreno, & Schweitzer, 2015; Brady, Wills, Jost, Tucker, & Van Bavel, 2017; Rimé, Finkenauer, Luminet, Zech, & Philippot, 1998). These cascades of emotions across a social network may play a key role in collective action because they contribute to a sense of identification, emphasize perceived injustice and empowerment, and in general increase the chance of engaging with others (Drury & Reicher, 2009; van der Linden, 2017; van Zomeren, Leach, & Spears, 2012).

However, emotions are fleeting processes that not only tend to spread but also tend to quickly decay (Garcia, Kappas, Küster, & Schweitzer, 2016; Gläscher & Adolphs, 2003; Kuppens, Oravecz, & Tuerlinckx, 2010). This means that after a short while, the intensity of emotions experienced by users decreases and the probability of emotional sharing and users' engagement is reduced (Christophe & Rimé, 1997; Garcia et al., 2016). The overall emotional intensity regarding a specific issue may be further reduced due to content that is either emotionally neutral or even opposite (such as positive emotions expressed in a negative context).

For emotional sharing to play a key role in social movements – despite the fact that emotions are fleeting and decay quickly – there must be emotional sharing processes that counter the natural decay in emotional intensity over time. What might these processes be? In the present work, we considered the possibility that emotional sharing in twitter replies may play a key role by shaping "appropriate" emotional responses. In particular, we examined situations that led to prolonged, increased emotional intensity, and tested whether when users responded to each other on social media, their replies expressed emotions that were even more aligned with the specific hashtag than the initial posts, thereby maintaining or even increasing emotional intensity.

Emotional Sharing on Social Media in the Context of Social movements

Social media are saturated with emotion-eliciting content. One reason social media are saturated with emotion-eliciting content is that attention is a key resource on social media, and eliciting emotion is a very good way to attract attention (Tufekci, 2013). A second reason social media are saturated with emotion is that in order to maximize revenue for social network companies, social network feeds are crafted to fit with users' previous likes and shares. This means that types of content that previously elicited emotions are repeated (Crockett, 2017).

When people encounter emotional content, they often share it with others (Kramer, Guillory, & Hancock, 2014). This can lead to cascades of emotional sharing processes (van der Linden, 2017). Two recent studies have powerfully made this point in the context of the 15M movement in Spain collective action (Alvarez et al., 2015), and in collective action associated with gun control and same-sex marriage (Brady et al., 2017). These studies point to the power of emotions to elicit social sharing, which might help fuel collective action.

Emotional Sharing in Original-Reply Pairs as a Vehicle for Social Movements

Potent as emotional sharing on social media may be, emotions tend to quickly decay and users then move to the next hashtag. What actions might counteract this decay? In answering this question, it is useful to note that in general, emotional sharing can occur via *original content*, by *replying*, or by *liking or sharing* certain content. Emotions expressed in original posts are often targeted towards a certain cause, person, or issue. Replies are often related to the same general issue as their corresponding original posts, but also reference the emotions expressed in the original post. Replying to certain posts and not others allows users to shape the emotions that are being expressed by emphasizing the appropriate emotions that fit the specific subject of discussion. Likes and shares do not involve producing original content on most social media platforms but rather are often used to support certain content produced by others.

How might these different forms of emotional sharing help maintain emotional intensity? In response to original tweets, users can write a reply that expresses emotions that are more aligned with the appropriate emotions regarding the issue at hand than those in the original. This may be especially relevant when replying to tweets that either express no emotions or different emotions than expected. For example, in a negative Twitter discourse regarding a case of police brutality that led to prolonged emotions, we would expect replies to be higher in negative intensity compared to their related original tweets. We would expect this effect to be driven mainly by replies to either neutral or positive emotions, because such tweets are those that contribute the most to overall reduction in negative emotions in this context. If indeed writing replies that are more negative in this context is desirable, we further expect that this increase in intensity would be rewarded on social media by more likes and shares, further perpetuating the motivation to express more negative emotion (see Crockett, 2017). Although analysis of social media cannot necessarily provide information regarding the motivations that drive stronger

emotional intensity in replies, previous research suggests that increased emotionality may be driven by users' desire to prove their dedication to the cause (Castells, 2012), to convince others to join (Goldenberg, Saguy, & Halperin, 2014) or to receive support (Crockett, 2017).

The Present Research

The goal of the present research was to examine emotional sharing on Twitter in the context of social movements. Our first hypothesis was that in a negative context, replies would be more negative than original tweets, and that this would be evident especially in response to neutral and positive tweets (Hypothesis 1). Our second hypothesis was that the larger the difference in negative emotional intensity between replies and originals, the more rewarded they would be by likes and shares (Hypothesis 2). We further expected that these effects would be reversed in a positive context, such that replies would be more positive than the original post (Hypothesis 3) and that a bigger positive difference would be more rewarded by likes and retweets (Hypothesis 4). To test our hypotheses, we conducted two Twitter studies. In Study 1, we examined the emotions expressed the context of the Ferguson unrest. In Study 2 we examined a positive context, looking at the celebration occurring in some quarters after Donald Trump's presidential victory.

Study 1:

Original Tweets and Replies in a Negative Context

The first goal of Study 1 was to examine whether replies are generally more negative than original tweets in a negative emotional context (Hypothesis 1). The second goal of Study 1 was to examine whether larger differences between replies and original tweets (i.e. more negative replies) would be rewarded by more likes and retweets (Hypothesis 2). In order to test

these two hypotheses, we analyzed the emotional content of tweet-reply pairs in the context of the Ferguson unrest that followed the shooting of Michael Brown on August 9th, 2014.

Method

Twitter. Twitter is a popular social network with millions of users around the world (Smith & Brenner, 2012). Unlike other social media platforms, 90% of all Twitter accounts are public (Takhteyev, Gruzd, & Wellman, 2012) which means that everyone can follow and read tweets (brief public messages posted to twitter.com) that are produced by these accounts. Twitter has played a crucial role in recent social movements and especially in the Ferguson unrest and the Black Lives Matter movement (Garza, 2015).

Data Collection. We gathered tweets in English which contained the keywords "#Ferguson", "#MichaelBrown", "#MikeBrown", "#Blacklivesmatter" and "#raceriotsUSA." Hashtags in Twitter identify discussion topics and create *ad hoc publics* in which like-minded individuals exchange messages in a given context (Bruns & Burgess, 2015). We chose to use hashtags as search terms based on recommendations from previous studies suggesting that hashtags provide a useful filter to receive both relevant content as well as content that is mostly produced by people who support the specific cause (Tufekci, 2014). Tweets were collected from a period of nearly four months starting from August 9th 12:00PM to December 8th 12:00AM. The data was downloaded via GNIP (gnip.com) which allows users to download full archives of tweets related to a certain search. The total number of collected tweets was 18,816,807 which were generated by 2,411,219 unique users. Out of these tweets, 3,149,026 were original tweets (users writing their own texts), 618,192 were replies (users replying to someone else's tweet), and 15,102,222 were retweets (users merely sharing previously written tweets). A small number of retweets included an original text and were therefore counted as both tweets and retweets.

Data Filtering. Our focus was on emotions expressed in replies. For this reason, we created tweet-reply pairs, in which the second tweet of the pair was always a reply to the first of the pair. Importantly, we made sure that the topic hashtags were used both in the reply and original of each pair to ensure that both replies and originals tweets were related to the specific context. Although this reduced the number of tweet-reply pairs, it ensured that both the writer of the original content and the user who replied were engaged in the same topic of discussion. In addition, we removed participants who wrote more than 20 tweets in order to avoid bots and news media. These efforts resulted in a sample of 255,164 tweet-reply pairs for the analysis.

Because GNIP archives tweets immediately after they are produced, we subsequently retrieved the data of each tweet at a later date to gather updated counts of retweets and likes through the Twitter Application Programming Interface (API). This effort took place on July 2016, two years after the incident. This was late enough to allow us to assume that the activity related to the tweets had fully decayed, since hashtags in Twitter stop attracting tweets very quickly (Wang, Ye, & Huberman, 2012). As some tweets were deleted between the GNIP archiving and our retrieval of the retweets and likes, our sample was reduced to 224,939 pairs of tweet-replies for the analysis. We compared the emotions expressed in the removed tweets to the larger sample and found no relevant differences.

One challenge in analyzing these specific Twitter datasets is how to detect the expected effects while taking into account statistical properties that may affect the outcomes. One major issue that we thought could pose a threat to our analysis was regression to the mean (for further discussion on regression to the mean on Twitter see Tomlinson, Bracewell, Krug, & Hinote, 2014). This is because if a certain tweet is evaluated as extremely negative, it is most likely that a reply to this tweet will be evaluated as lesser in intensity. This also means that if a certain dataset

has more negative than positive tweets, regression to the mean may lead replies to be more positive than originals. This consideration motivated us to create more balanced datasets of original tweets, in which the number of negative, neutral, and positive original tweets is equal in both studies. Having such a balanced data set could reveal differences that occur over and above the effects of regression to the mean.

Because the Ferguson unrest was a context in which most of the emotions expressed were negative, the number of negative, neutral, and positive original tweets was not balanced and was skewed towards negative original tweets (negative emotions = 107,454; neutral = 88,482; positive = 29,003). In order to overcome this bias and to reduce the effects of regression to the mean, we created a balanced subsample with equal numbers of negative, neutral, and positive tweets (N = 87,009 tweets). This balanced sample was built using the full number of positive tweets (the smallest amount) and randomly sampling an equal number of neutral and negative tweets. We used this balanced sample to test for differences between the reply and original tweet.

We quantified an approximation of the positive and negative emotions expressed in tweets using the sentiment analysis tool **SentiStrength**. SentiStrength is a specifically designed sentiment analysis tool for short, informal messages from social media (Thelwall, Buckley, & Paltoglou, 2012). SentiStrength takes into account syntactic rules like negation, amplification, and reduction, and detects repetition of letters and exclamation points as amplifiers. Compared to other sentiment analysis tools in standardized benchmarks, SentiStrength has been shown to be among the state-of-the-art sentiment analysis methods for short social media posts, and Twitter in particular (Ribeiro, Araújo, Gonçalves, André Gonçalves, & Benevenuto, 2016).

The output produced by SentiStrength is based on a bivariate model of emotions (Larsen & Diener, 1992) which measures emotions on two separate dimensions, positive intensity and

negative intensity. For any given text, SentiStrength produces two scores, one score for positive intensity (ranging from 1 to 5) and one score for negative intensity (ranging from -1 to -5). We added the negative and positive scores in order to create one scale of emotional intensity ranging from -4 to 4, -4 indicating high intensity negative emotions and 4 indicating high intensity positive emotions.

Data Analysis. Using our balanced dataset, what outcomes would constitute support for our hypotheses? One analysis would involve looking at replies to neutral tweets. For example, negative replies to neutral tweets may reveal a negative bias in the tendency to reply. A second analysis would involve comparing the difference between replies and originals for positive and negative original tweets. If only regression to the mean is at play, then the difference between replies and originals should be equal for positive and negative original tweets. However, if the comparison of the difference between replies and originals for positive and negative tweets is not equal, this would provide an indication for a bias in replies. Finally, as mentioned in the second and fourth hypotheses, a difference in the number of likes and retweets as a function of the difference in emotional intensity between replies and originals would provide a third indication that such a difference has an important meaning for the way emotional sharing occurs on social media.

Results

Difference Between Replies and Original Tweets. We first compared the overall difference in emotional intensity within the tweet-reply pairs. This was done by conducting a mixed model analysis using the type of tweet (original versus reply) as the independent variable and the degree of emotion expressed as a dependent variable. In addition, we used a by-tweet random variable to make sure that comparisons are made within each tweet-reply pair.

Supporting our first hypothesis, results suggested that overall, replies were significantly more negative than original tweets (*b* = -.*253* [-.261, -.245], *SE* = .04, *t*(115,853) = -60.55, *p* < .001, *d* = -.22, R^2 = .56, Figure 1). Cohen's d for repeated measures was calculated based on recommendations by Westfall and colleagues (Westfall, Kenny, & Judd, 2014). R^2 calculation was conducted based on recommendation by Xu (2003).

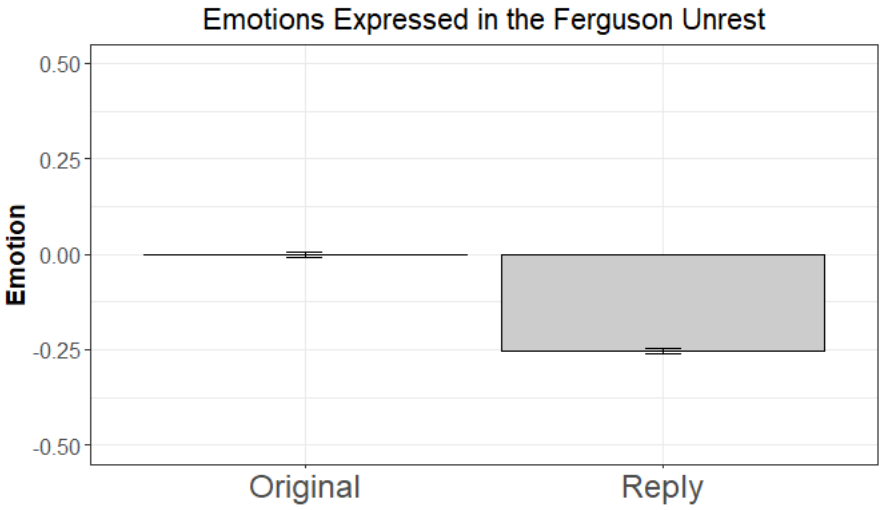

Figure 1. Mean emotions expressed in replies and original tweets in Study 1 (Ferguson unrest). The mean of original tweets is equal to zero due to the fact that the dataset was balanced to ensure an equal number of positive, neutral and negative original tweets. Error bars are 95% confidence intervals.

We further examined whether this difference was driven by replies to neutral positive or negative tweets. We therefore compared the difference between the reply and original tweets for neutral, positive, and negative tweets. Our analysis suggested that for neutral tweets, replies were more negative than original tweets (*b* = -.*226* [-.238, -.213], *SE* = .006, *t*(171,855) = -62.30, *p* < .001, *d* = -.22, R^2 =.10). This was also true for positive tweets (*b* = -.*965*[-.979, -.950], *SE* = .007, *t*(205,949) = -233.88, *p* < .001, *d* = -.85, R^2 = .43). Finally, for negative tweets, replies were more positive (*b* = .*431* [.418,.444], *SE* = .006, *t*(260,547) = 116.73, *p* < .001, *d* = .40, R^2=.30).

Importantly, the effects in both the negative and positive conditions might have been influenced by regression to the mean, leading to the fact that negative Tweets were more likely to receive more positive replies and positive Tweets received more negative replies. We therefore compared the difference in emotion between replies and original tweets across these two conditions (negative original tweets and positive original tweets). This was done by reverse scoring the difference between replies and originals for positive originals and then comparing it to the difference for negative originals (equivalent to comparing the sizes of the red lines in Figure 2). The difference between emotions in originals and replies for positive original tweets was significantly larger than the difference for negative tweets (b = -.59 [-.613, -.568], SE = .01, $t(34{,}352)$ = -51.42, p < .001, d =-.48, R^2=.49). These findings suggest that the negative difference between replies and original tweets was mostly influenced by users responding in a more negative way to positive and neutral tweets.

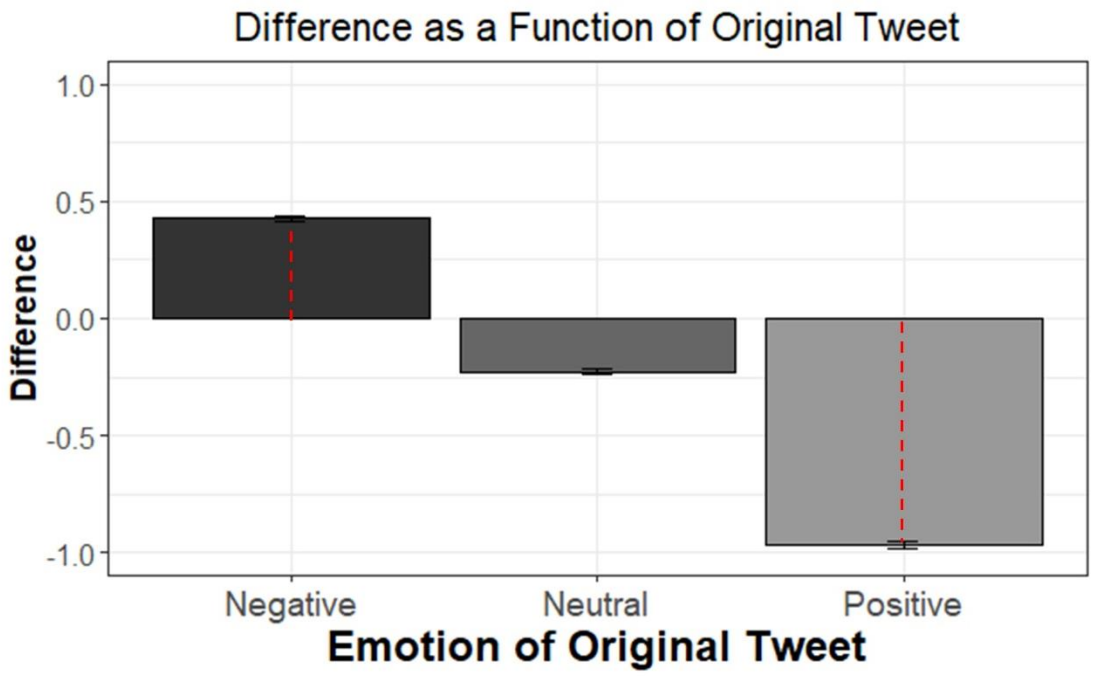

Figure 2. Mean differences between reply and original as a function of the emotions expressed in the original tweet. For negative tweets, replies are more positive than originals. For neutral and positive original tweets, replies are more negative than originals. The difference between replies

and originals for positive tweets was significantly larger than the difference for negative tweets. Error bars are 95% confidence intervals.

Likes and Retweets. To test whether more negative replies were rewarded more on social media, we examined whether the emotion difference between replies and original tweets predicted the number of likes and retweets that Twitter users gave to those replies. As the distribution of the number of likes and retweets per tweet was skewed toward zero, we applied a log-modulus transformation to the counts of retweets and likes of each tweet. We then conducted a mixed model analysis using the emotion difference between replies and original tweets to predict the number of likes and retweets of the reply tweet. Similar to our previous analysis, we used a by-tweet random variable for our analysis. Looking first at likes, results suggested that the more negative replies were relative to original tweets, the higher number of likes (b = -.*004*[-.005, -.002], SE = .001, t(215,453) = -4.68, p < .001, R^2 = .16) (see Figure 3A). The same effect was found when analyzing retweet counts (b = -.*006*[-.008, -.005], SE = .001, t(217,302) = -7.20, p < .001, R^2 = .19) (see Figure 3B). These findings support our second hypothesis that online social networks reward negative difference between replies and originals.

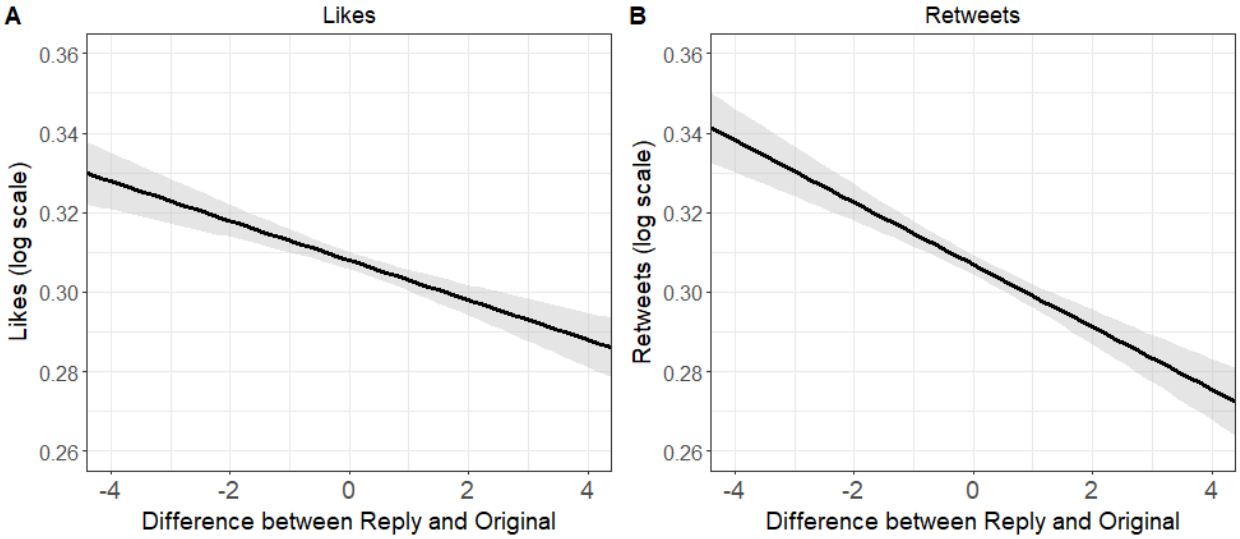

Figure 3. Difference between replies and originals predicting the number of (log transformed) likes (Panel A) and retweets (Panel B) in tweets related to the Ferguson unrest.

The findings of Study 1 point to a tendency to reply to tweets with more negative tweets in a context that was considered negative to the relevant audience. This tendency is mainly driven by course correction of positive and neutral tweets. Furthermore, our findings suggest that greater increase in difference between replies and original tweets is rewarded by more likes and retweets. One question, however, is whether these effects are restricted to negative contexts, or whether they could also occur in a positive context, such that replied would be more positive than original tweets. The purpose of Study 2 was to examine this question.

Study 2:

Original Tweets and Replies in a Positive Context

The goals of Study 2 were to see whether replies are generally more positive than original tweets in a positive context (Hypothesis 3) and whether social media generally reward replies that are more positive than original tweets by liking and retweeting them more (Hypothesis 4). In order to achieve these goals, we analyzed the emotional content of tweet-reply pairs in tweets that celebrated the victory of Donald Trump as president of the United States in November 2016.

Method

Data Collection. We gathered tweets in English which contained the keywords "#PresidentTrump", "#TrumpWinner", "#MakeAmericaGreatAgain", "#MAGA", "#TrumpWon and "#Trump2020." As in Study 2, we chose to use hashtags as search terms based on recommendations from previous studies (Barberá et al., 2015; Tufekci, 2014). Tweets were collected from a period of two months starting from November 9th 12:00 AM to January 8th 11:59AM. The difference in length from Study 1 was causes by a change in Twitter's policy

regarding data downloading. The total number of collected tweets was 8,417,356 which were generated by 865,136 unique users. Out of these tweets, 1,520,290 were original tweets (users writing their own texts), 321,358 were replies (users replying to someone else's tweet), and 6,579,781 were retweets (users merely sharing previously written tweets). A small number of retweets included an original text from the user and were therefore counted as both tweets and retweets. As in Study 1, we used the sentiment analysis tool SentiStrength, which is specifically designed for short, informal messages from social media (Thelwall et al., 2012).

Data Filtering. We used a data filtering protocol that was identical to that of Study 1, leading to a sample of 23,731 pairs for the analysis. Because GNIP archives tweets immediately after they are produced, we subsequently retrieved the data of each tweet at a later date to gather updated counts of retweets and likes through the Twitter Application Programming Interface (API). This was done on June 2017, six months after the elections in the United States. As some tweets were deleted between the GNIP archive and our retrieval of the counts of retweets and likes, our sample was reduced to 19,306 pairs of tweet-replies for the analysis. We compared the emotions expressed in removed tweets to the larger sample and found no relevant differences. The number of negative, neutral, and positive original tweets was not balanced and was biased towards positive original tweets (negative emotions = 5,962; neutral = 6,441, positive = 6,902). In order to overcome this bias, we created a balanced dataset that sampled from the negative, neutral, and positive tweets in order to create an equal number of negative, neutral, and positive originals (N = 16,661 tweets). This was done in the same way as in Study 1.

Results

Difference between Reply and Original Tweet. We compared the difference in emotion within the tweet-reply pairs by conducting a mixed model analysis using the type of tweet

(original versus reply) as the independent variable and the degree of emotion expressed as the dependent variable. In addition, we used a by-tweet random variable to make sure that comparisons were made within each tweet-reply pair. Results suggested that overall replies were significantly more positive than original tweets (*b* =*.114* [.092, .137], *SE* = .01, *t*(23,706) = 9.97, $p < .001$, *d* = .10, $R^2 = .44$, Figure 4).

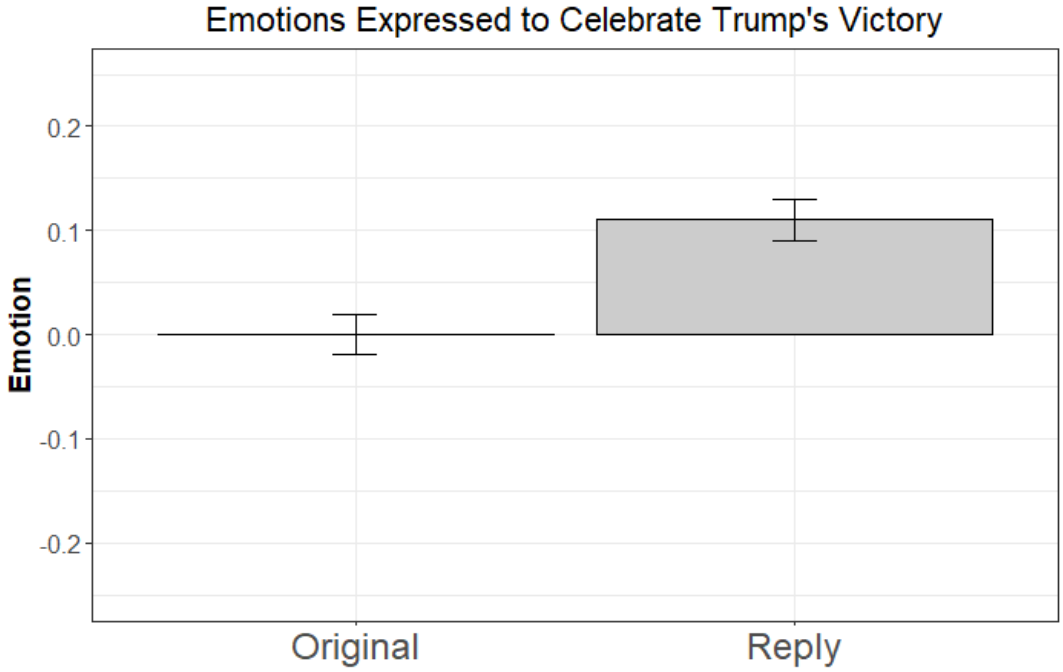

Figure 4. Mean emotions expressed in replies and original tweets in Study 2 (Celebrating Trump's victory). The mean of original tweets is equal to zero due to the fact that the dataset was balanced to ensure an equal number of positive, neutral and negative original tweets. Error bars are 95% confidence intervals.

To examine whether this difference was driven by replies to negative, neutral or positive tweets, we compared the difference between the emotions in reply and original tweets for negative, neutral, and positive tweets. Our analysis suggested that for neutral tweets, replies were more positive than original tweets (*b* =*.132* [.110, .164], *SE* = .01, *t*(59,164) = 14.23, $p < .001$, *d* = .13, $R^2 = .12$). This was also true for negative tweets (*b* =*1.21* [1.17, 1.25], *SE* = .01, *t*(47,489)

= 114.67, $p < .001$, $d = 1.06$, $R^2 = .37$). Finally, for positive tweets, replies were more negative (*b* =*-1.00* [1.03, -.96], *SE* = .01, $t(52{,}812) = -101.33$, $p < .001$, $d = -.89$, $R^2 = .39$).

The effects in both the negative and positive conditions might have been influenced by regression to the mean, meaning that negative tweets were more likely to receive more positive replies and positive tweets were more likely to receive more negative replies. We therefore compared the difference in emotion between replies and original tweets across these two conditions (equivalent to comparing the red lines in Figure 5). This was done by reverse scoring the difference between replies and originals for positive originals and then comparing it to the difference for negative originals. Results suggested that the difference between emotions in originals and replies for negative original tweets was significantly larger than the difference for positive tweets (*b* = *.13* [.073, .192], *SE* = .03, $t(6{,}407) = 4.37$, $p < .001$, $d = .10$, $R^2 = .36$). These findings suggest that the positive difference between replies and original tweets was mostly influenced by users responding in a more negative way to positive and neutral tweets.

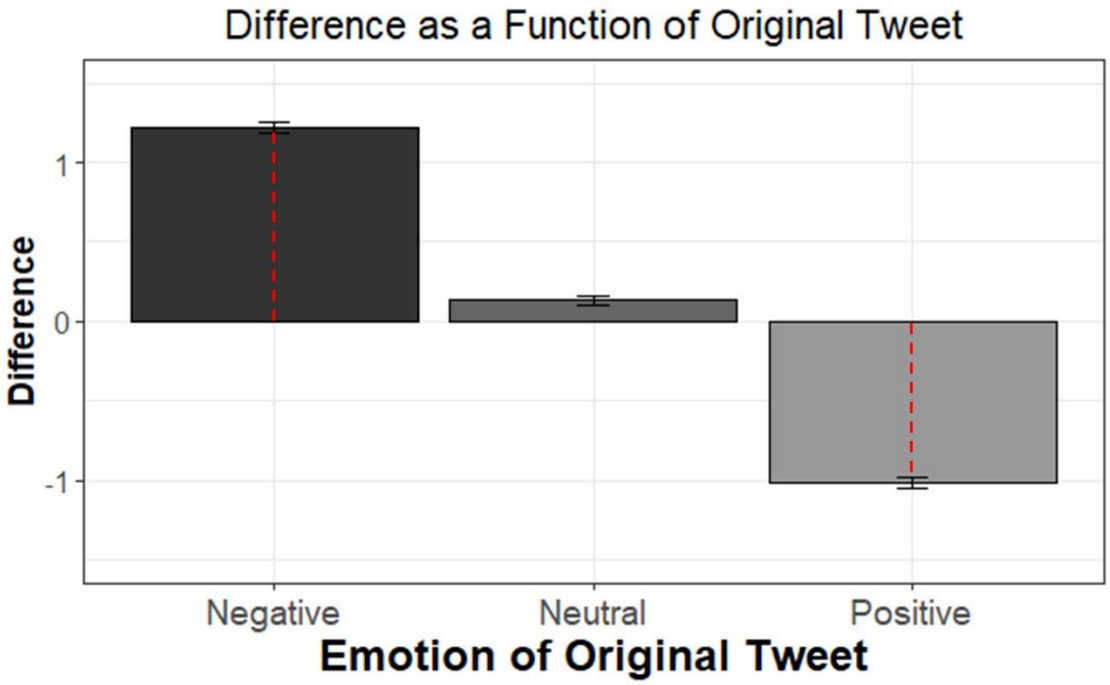

Figure 5. Mean differences between reply and original as a function of the emotions expressed in the original tweet. For negative tweets and neutral tweets, replies are more positive than originals. For positive original tweets, replies are more negative than originals. The difference between replies and originals for negative tweets was significantly larger than the difference for negative tweets. Error bars are 95% confidence intervals.

Likes and Retweets. We examined whether the difference between replies and original tweets predicted the number of likes and retweets that were received to these replies. As in Study 1, we applied a log-modulus transformation of the counts of likes and retweets. Looking first at likes, results suggested that the more positive replies were relative to original tweets, the higher amount of likes (b =*.010*[.004, .016], SE = .003, t(18,575) = 3.317, $p < .001$, R^2 = .46) (see Figure 6A). The same effect was found when looking at retweets, such that increase in negative escalation lead to an increase in the amount of retweets (b =*.005*[.0004, .010], SE = .002, t(15,154) = 2.17, $p < .001$, R^2 = .09) (see Figure 6B). These findings suggest that replies that were more positive than original tweets received both more likes and retweets.

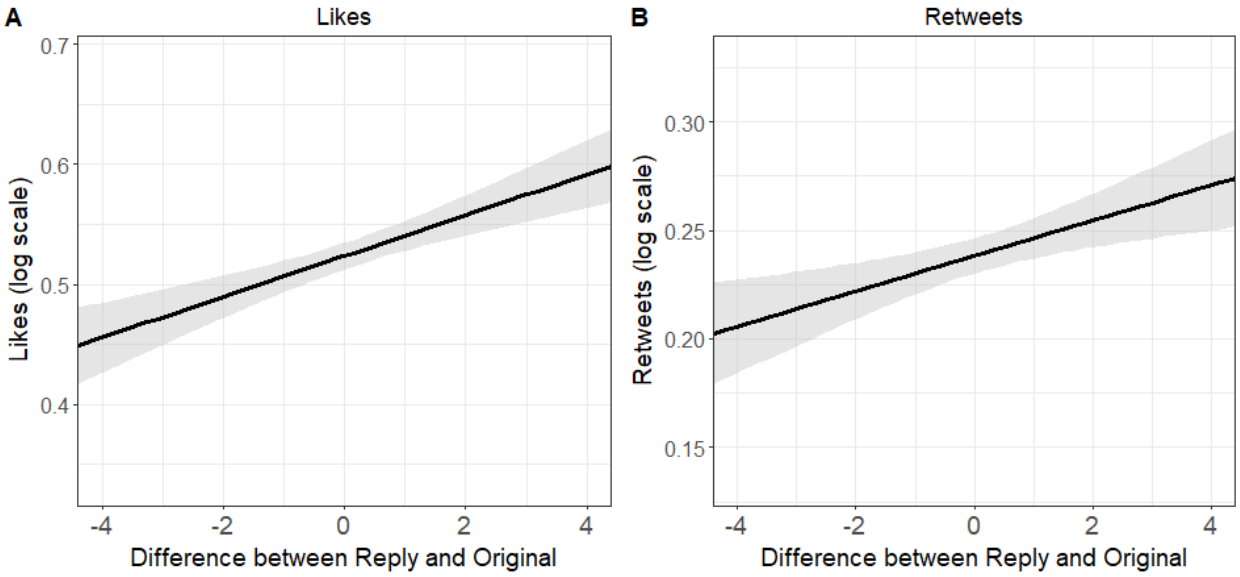

Figure 6. Difference between replies and originals predicting the number of (log transformed) likes (Panel A) and retweets (Panel B) in tweets celebrating Trump's victory.

The findings of Study 2 are a mirror image of those revealed in Study 1. Results point to a tendency for replies to be more positive than their corresponding original tweets. This tendency is mainly driven by replies to neutral and negative tweets. Furthermore, our findings suggest that higher positive difference between replies and original tweets is rewarded by more likes and retweets.

General Discussion

Social media is thought to play a key role in social movements. However, it is not yet clear how social dynamics contribute to the increased levels of emotional intensity that support social movements. We hypothesized that, in these unique cases in which social movements emerges, social media replies might express emotions that are more aligned with a particular movement than the original posts. The predicted differences in emotions were found both in a negative context, looking at tweet-reply pairs related to the Ferguson unrest, and in a positive context, looking at tweet-reply pairs celebrating Donald Trump's election victory. Furthermore, our findings show that social media rewarded a bigger difference in emotional intensity between a reply and original tweet by retweeting and liking these tweets.

Increased Emotional Intensity as a Driver of Collective Action

Strong emotions are an important factor in fueling collective action. Emotions not only motivate individuals to action but also amplify other psychological processes that promote collective action such as increased group identification (Drury & Reicher, 2009; Tajfel & Turner, 1979; van Zomeren, Spears, Fischer, & Leach, 2004), increased perceived efficacy and empowerment (Drury & Reicher, 2005, 2009) and a clearer perception of the right course of action (McGarty et al., 2014). Examining situations that lead to increased emotional intensity may be a useful gateway to identifying the mechanisms propelling groups to action.

One question is whether increased emotionality on social media can contribute to collective action that occurs *outside of social media*. The answer is of course yes, to some extent, but recent work suggests that online collective action may be less effective than expected. For example, an analysis of the outcomes of the "ALS Ice Bucket Challenge", showed that although the campaign led millions of people to upload videos of themselves pouring an ice on their heads, only one in five mentioned a donation to ALS (van der Linden, 2017). Similar results were seen in the "Save Darfur" campaign (Lewis, Gray, & Meierhenrich, 2014). Further work therefore should be done not only to examine the role of emotions on social media in promoting collective action outside of social media.

Increased Emotional Intensity as a Mechanism for Polarization

The processes described in the present studies may not only lead to the maintenance of emotional intensity but may also lead to increased conflict and polarization. Users who express emotions that contradict the general sentiment of a certain movement may feel rejected when responded to with strong emotions (Szczurek, Monin, & Gross, 2012). This effect may contribute to the formation of echo chambers, in which people with similar opinions and emotions are connected to each other but not to people who don't share their views and emotions (Boutyline & Willer, 2017; Brady et al., 2017; Sterling & Jost, 2017). The current study builds on prior findings by proposing one mechanism for their occurrence.

Once echo chambers are formed on social media, increased emotional intensity in replies may be used by users to indicate their dedication to a certain cause. The fact that users are rewarded by others when writing replies that are more emotional than original tweets provides an initial indication for a reward mechanism that helps sustain strong emotions (Crockett, 2017). Over time, such emotional dynamics may play a role in further polarization.

Limitations and Future Directions

The two studies presented here provide the first test of the tendency for increased emotional intensity in replies. However, several important limitations should be noted.

First, looking at actual tweets does not yet answer questions regarding the specific motivations that drive the processes that were observed in our study. One possibility is that users attempt to maximize their outrage in order to prove their dedication to the cause (Alvarez et al., 2015; Castells, 2012). A second possibility is that when participants are expressing stronger emotions they express stronger emotions in order to keep other users emotionally involved (Goldenberg, Halperin, van Zomeren, & Gross, 2016; Goldenberg et al., 2014). A third possibility is that users are merely maximizing their emotions to receive more support from others (Crockett, 2017). Testing these ideas in carefully controlled experiments is important to the understanding of these processes.

A second limitation of the current work is that we do not know how much of the difference between replies and tweets is driven by a selection bias in which users who write original tweets are different in some way than those who write replies who may have a lower emotional threshold. Further work should examine this question by comparing users who tend to write originals and replies.

Overall, social media provides us with exciting opportunities to understand social processes. While Twitter datasets do not provide all the information required to understand users' psychology, they can be an important tool to learn about processes that can later be examined in controlled experiments.

References

Alvarez, R., Garcia, D., Moreno, Y., & Schweitzer, F. (2015). Sentiment cascades in the 15M movement. *EPJ Data Science, 4*(1), 1–13.

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The critical periphery in the growth of social protests. *PLoS ONE, 10*(11), 1–15. https://doi.org/10.1371/journal.pone.0143611

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology, 38*(3), 551–569. https://doi.org/10.1111/pops.12337

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bruns, A., & Burgess, J. (2015). Quantitative approaches to comparing communication patterns on Twitter. In N. Rambukkana (Ed.), *Hashtag publics: The power and politics of discursive networks* (pp. 13–28). New York, NY: Peter Lang. https://doi.org/10.1080/15228835.2012.744249

Castells, M. (2012). *Networks of outrage and hope : Social movements in the internet age* (2nd ed.). Cambridge, UK: Polity Press.

Christophe, V., & Rimé, B. (1997). Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology, 27*(1), 37–54.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*, 769–771.

Drury, J., & Reicher, S. (2005). Explaining enduring empowerment: A comparative study of collective action and psychological outcomes. *European Journal of Social Psychology, 35*(1), 35–58.

Drury, J., & Reicher, S. (2009). Collective psychological empowerment as a model of social change: Researching crowds and power. *Journal of Social Issues, 65*(4), 707–725.

Garcia, D., Kappas, A., Küster, D., & Schweitzer, F. (2016). The dynamics of emotions in online interaction. *Royal Society Open Science*. https://doi.org/10.1098/rsos.160059

Garza, A. (2015). A herstory of the #BlackLivesMatter movement. *ProudFlesh: New Afrikan Journal of Culture, Politics and Conscioousness*.

Gläscher, J., & Adolphs, R. (2003). Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *Journal of Neuroscience, 23*(32), 10274–10282.

Goldenberg, A., Halperin, E., van Zomeren, M., & Gross, J. J. (2016). The process model of group-based emotion: Integrating intergroupe emotion and emotion regulation perspectives. *Personality and Social Psychology Review, 20*(2), 118–141. https://doi.org/10.1177/1088868315581263

Goldenberg, A., Saguy, T., & Halperin, E. (2014). How group-based emotions are shaped by

collective emotions: Evidence for emotional transfer and emotional burden. *Journal of Personality and Social Psychology*, *107*(4), 581–596. https://doi.org/10.1037/a0037462

Juris, J. S. (2012). Reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist*, *39*(2), 259–279. https://doi.org/10.1111/j.1548-1425.2012.01362.x

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. *Review of Personality and Social Psychology*, *13*, 25–59. https://doi.org/10.1093/scan/nsl006

Lewis, K., Gray, K., & Meierhenrich, J. (2014). The structure of online activism. https://doi.org/10.15195/v1.a1

McGarty, C., Thomas, E. F., Lala, G., Smith, L. G. E., & Bliuc, A.-M. (2014). New technologies, new identities, and the growth of mass opposition in the Arab Spring. *Political Psychology*, *35*(6), 725–740. https://doi.org/10.1111/pops.12060

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench: A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

Rimé, B., Finkenauer, C., Luminet, O., Zech, E., & Philippot, P. (1998). Social sharing of emotion: New evidence and new questions. *European Review of Social Psychology*, *9*(1), 145–189. https://doi.org/10.1080/14792779843000072

Smith, A., & Brenner, J. (2012). *Twitter Use 2012*. Washington, DC. Retrieved from http://www.looooker.com/wp-content/uploads/2013/05/PIP_Twitter_Use_2012.pdf

Smith, L. G. E., Gavin, J., & Sharp, E. (2015). Social identity formation during the emergence of the occupy movement. *European Journal of Social Psychology*, *45*(7), 818–832. https://doi.org/10.1002/ejsp.2150

Sterling, J., & Jost, J. T. (2017). Moral discourse in the Twitterverse. *Journal of Language and Politics,* 1–27. https://doi.org/10.1075/jlp.17034.ste

Szczurek, L., Monin, B., & Gross, J. J. (2012). The stranger effect: The rejection of affective deviants. *Psychological Science*, *23*(10), 1105–1111. https://doi.org/10.1177/0956797612445314

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 33–47.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social*

Networks, *34*(1), 73–81. https://doi.org/10.1016/J.SOCNET.2011.05.006

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Tomlinson, M. T., Bracewell, D. B., Krug, W., & Hinote, D. (2014). #impressme: The language of motivation in user generated content. In A. Gelbukh (Ed.), *Computetional linguistics and intelligent text processing.* (pp. 176–187). Heidelberg, Germany: Springer.

Tufekci, Z. (2013). "Not this one": Social movements, the attention economy, and microcelebrity networked activism. *American Behavioral Scientist*, *57*(7), 848–870.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 505–514.

Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, *62*(2), 363–379. https://doi.org/10.1111/j.1460-2466.2012.01629.x

van der Linden, S. (2017). The nature of viral altruism and how to make it stick. *Nature Human Behaviour*, *1*, 1–3. https://doi.org/10.1038/s41562-016-0041

van Zomeren, M., Leach, C. W., & Spears, R. (2012). Protesters as "passionate economists": A dynamic dual pathway model of approach coping with collective disadvantage. *Personality and Social Psychology Review*, *16*(2), 180–199. https://doi.org/10.1177/1088868311430835

van Zomeren, M., Spears, R., Fischer, A. H., & Leach, C. W. (2004). Put your money where your mouth is! Explaining collective action tendencies through group-based anger and group efficacy. *Journal of Personality and Social Psychology*, *87*(5), 649–664.

Wang, C., Ye, M., & Huberman, B. A. (2012). From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 244–252). Beijing, China. Retrieved from http://dl.acm.org/citation.cfm?doid=2339530.2339573

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, *22*(22), 3527–3541.