A Appendix

This appendix contains several additional analyses and figures as well as further details on the coding of the independent and dependent variables.

A.1 Further Details on Coding of Variables

A.1.1 Coding of Reputational Favorability

Data were coded in three stages. First, I selected a sample of 1,000 replies for crowdsourced manual coding.[1] Crowdsourced coding – human coding by the online "crowd" on such platforms as Mechanical Turk and Crowdflower – provides a rapid and cost-effective method for manually coding data that is becoming increasingly popular in management and accounting research (Grenier, Lowe, Reffett, & Warne, 2015; Rennekamp, 2012). These 1,000 tweets were uploaded to Crowdflower (www.crowdflower.com). On the site I created a set of instructions for the coders and manually coded 55 randomly selected replies as having either a negative, neutral, or positive sentiment; I also provided a brief justification for each score that was visible to the coders as both an initial practice set and for providing intermittent feedback as they proceed with coding. Figure A.1 contains a screenshot of the instructions given to Crowdflower workers to code the sentiment in public replies, along with a sample tweet.

[Insert Figure A.1 here]

Each of the 1,000 tweets was coded by at least three Crowdflower coders. The majority of tweets coded had 100% agreement across the three coders. Less than 1% (n=9) saw complete disagreement across the three coders, while roughly 30% saw two of three coders agree on sentiment ratings. From these values Crowdflower returns a "best score," which factors in the level of confidence in each of the three coders (based on past performance); however, for the most part the "best score" is the majority vote in each case. Agreement with this crowdsourced score and my own manual codes was 89% with a Cohen's kappa score of 0.817 ($\kappa = 0.817$), indicating a high level of agreement.[2]

I then proceeded to implement a machine learning technique to code the remainder of the replies. Machine learning approaches differ from automated, unsupervised, lexicon-based approaches used in prior CSR research in that they are supervised techniques, requiring the research to make decisions and to "train" the machine learning algorithm. There are several steps involved. In these techniques the researcher takes a sample of the data, codes it, and then trains the machine learning algorithm by fine tuning select parameters. In the present study, first the 1,000 manually coded tweets were divided into training (85% of cases) and test datasets (15% of cases). Tweets were then pre-processed for machine learning by making content lowercase, by "stemming" words (e.g., "making" becomes "make") and by removing "stop words" (common words such as "the").

[1] Specifically, I selected the 304 replies to the 1,000 randomly selected firm tweets, plus an additional 696 randomly selected replies to firm tweets, for a total of 1,000 replies.

[2] Commonly cited guidelines by Landis and Koch (1977) refer to kappa scores of .81 or greater as "almost perfect."

Three popular machine learning algorithms – Naïve Bayes, Decision Trees, and Support Vector Machines (SVM) – were trained. Each algorithm was trained separately by modifying key algorithm parameters (e.g., by selecting how many text features to consider in the algorithm)[3] and, after each round of training, assessing the classifier's accuracy by comparing the algorithm scores generated on the training dataset to those in the test dataset. In line with prior research on tweets (Go et al., 2009), the SVM algorithm achieved the highest accuracy. Nave Bayes achieved 75.8% accuracy and the Decision Tree algorithm achieved 76.9% accuracy, both somewhat lower than initial performance with SVM. Consequently, SVM was chosen and parameters fine-tuned until the highest level of accuracy was achieved.

With the classifier fully trained, agreement on the 3-code sentiment variable (-1, 0, +1) between the manual coding and SVM coding was 81.3%. This fits well with expectations based on prior research that has found accuracy in sentiment classification to be around 82% (e.g., Go et al., 2009 achieved 82.2% accuracy). Accuracy is even higher with three binary variables derived from the sentiment variable: 89.0% for positive, 80.2% for neutral, and 94.5% for negative.

With the classification algorithm trained and tested, it was then used to generate sentiment scores for the 4,247 remaining replies. For comparison, scores were also created for the 1,000 replies that were manually coded. The agreement between the 1,000 SVM-coded variables and the 1,000 manually coded negative, neutral, and positive binary variables is 93.3%, 79.6%, and 86.1%, respectively.[4] This indicates a high level of inter-coder reliability. As a further check, I also conducted several inter-coder reliability scores. Cohen?s kappa values were from a low of 0.5827 for neutral, considered "moderate agreement" (Landis & Koch, 1977) 0.604 for negative, and 0.677 for positive, both considered "substantial agreement" (Landis & Koch, 1977).

In short, through the manual coding and supervised machine learning process each of the 5,247 replies received a value of "1" on one of the following three binary variables derived from the three-value (-1, 0, +1) sentiment variable ? positive, neutral, and negative. Recall that the analysis is a message-level analysis, with the messages being the 18,722 company tweets. In a final step, the reply data were therefore merged into the company tweet dataset. Because some firm tweets could receive more than one reply, the reply-level dataset (n=5,247) was collapsed by the company tweet that was the target of the response and merged into the company tweet database (n=18,722), in the process creating the final two binary variables for analysis: Positive Reply, with a value of "1" indicating firm tweets that receive at least one positive reply; and Negative Reply, with a value of "1" indicating firm tweets that receive a negative reply.

[3]For instance, in training the data, the highest accuracy came with choosing 10% of the features for the algorithm.

[4]I then ran these 1,000 tweets against the unsupervised/automated lexicon-based approach ANEW (Nielsen, 2011) and saw 64% agreement with the manually coded Crowdflower codes. Splitting the sentiment variable into negative, neutral, and binary variables the agreement with the hand-coded tweets is better at 89.5% (negative), 67% (neutral), and 71.5% (positive), yet still underperforms the SVM-based classifier by roughly 4%, 13%, and 14.5% on the negative, neutral, and positive variables, respectively.

A.1.2 Coding of Communication Tactics

As with reply sentiment, a machine learning classifier was trained to code the 17,222 non-hand-coded firm tweets. To improve the accuracy of the classifier only hand-coded tweets with agreement by at least 2 of 3 Crowdflower coders were used (1,257 of the 1,500 tweets). The final classifier was an SVM algorithm that achieved 80.0% overall accuracy compared to the test data set (comprising 12.5% of the 1,500 manually coded tweet); specifically, agreement with the hand-coded tweets was 81.4% with disclosure, 86.9% with public education, 97.2% with marketing, and 93.8% with no information. Given the high levels of accuracy the classifier was then used to classify the remaining tweets (i.e., assign one of the four values on these variables).

As an additional check on the reliability of the coding, scores from this round of machine learning were also compared to the 1,500 hand-coded tweets. For the four-category variable, accuracy between the 1,500 hand-coded tweets and the machine learning-coded tweets was 82.53% ($\kappa = 0.664$). For disclosure it was 82.53%, for public education it was 88.27%, for marketing it was 98.8%, and for non-informational it was 94.67%. Moreover, comparing the machine learning-coded tweets to the 200 initial hand-coded tweets, for the four-category variable agreement was 82.5%. For the four binary variables derived from this variable and used for analyses, agreement for Disclosure was 84.5% with a kappa value of 0.687, for Public Education it was 89% with a kappa value of 0.703, for Marketing it was 99% with a kappa score of 0.911, and for non-informational it was 92.5% with a kappa value of 0.702. In short, all would be considered at least "substantial agreement" by Landis and Koch (1977).

The other five (non-informational) tactics, however, could be coded through a different, more automated process. Namely, I developed custom algorithms for coding each of the remaining 18,522 tweets for the existence of the five tactics. For instance, one of the tactics coded was the *politician mention*, where firms made an effort to thank, congratulate, or interact with politicians. To find mentions of politicians, a list of all Twitter users mentioned in the 18,722 tweets that had usernames starting with "@Gov," "@Sen" "@Mayor," and "@Rep," such as @RepJoeKennedy and @SenGillibrand, was compiled then verified to ensure that these were Twitter accounts of politicians. Other politicians that were discovered to be mentioned, such as @NancyPelosi, were also added to the list. In the end a list of 158 verified politicians was created; at that stage automated Python code was written to identify every tweet among the 18,722 firm messages that contained one or more mentions of users from the list of politicians. Similar algorithms were written for the other tactics categories and were refined until coding accuracy was well above 90% compared to hand coding.

A.2 Account-Level Analyses

The paper is built around message-level analyses. However, some may be interested in seeing some relevant account-level analyses. Figure A.2 contains the aggregate number of retweets and positive replies garnered by each of the accounts over the course of 2014.

[Insert Figure A.2 here]

Figure A.3, meanwhile, shows the number of new followers acquired over the course of 2014. While the number of followers was not a key variable in the analyses, it is an important

indicator of the influence of a Twitter account.

[Insert Figure A.3 here]

What is interesting about these figures is, first, how the distribution appears to follow a power law distribution. Second, the clear "winner" in all three figures, *BofA_Community*, spent a lot of money on its various CSR campaigns over the course of the year.

A.3 Actor-Network Analyses

Figure A.4 contains a depiction of a network analysis inspired by Bruno Latour's Actor-Network Theory (A.N.T.). The figure depicts a network of ties among the actors, the tactics, and the audience reactions in the 18,722 firm tweets. Specifically, for each company tweet data were available on who sent it, which tactics were employed, and the audience reactions (retweeted, positive reply, neutral reply, negative reply, and none/ignored). Accordingly, I looped over each tweet and created *edges* based on the combination of sender, tactics, and reactions. I ran this in NetworkX in Python, exported to Gephi, and laid out the network using a "Force Atlas" layout. Edge line thickness reflects weight and node size reflects degree.

[Insert Figure A.4 here]

A brief examination of this figure tells us the following. First, at the center of the network are three tactics – *topic ties*, *user mentions*, and *disclosure* – as well as two outcomes – *ignored* and *retweet* – as well as a handful of Twitter accounts. We can also look at individual accounts and see which tactics or outcomes they are closer to, or look at individual tactics and see which outcomes they are closer to, etc. While I ultimately decided not to analyze these data in depth, it remains a potentially fruitful avenue for future research.

A.4 Robustness Tests

A.4.1 Additional Measures of Firm Size

Number of Employees. I also conducted a number of additional analyses (not shown) to verify the robustness of the above tests. First, I incorporated alternative measures to tap the size of the firm. The argument implicitly made above is that it the number of followers of the firm is the more proximate, and hence more important, measure of company size (i.e., firms with greater assets or revenues or more employees will generally be associated with higher numbers of followers). Nevertheless, as a check on this assumption I also ran the above six models with number of employees as a control. First, when included in the three regressions with the aggregate tactics Information, Interaction, and Tie-building, the number of employees is significantly and positively associated with Retweet in model 1, and is not significant in model 3 (DV = Positive Reply) or model 5 (DV = Negative Reply). In models 1, 3, and 5 there are no changes in sign or significance for any of the other independent or control variables. In model 2, meanwhile, the number of employees does not obtain significance and there are no changes in sign for any of the other variables, but two previously non-significant variables obtain significance: Mobilization is positively associated with Retweet

(p=.082) and Politician Mention is negatively associated (p<.01). Similarly, in model 4 (DV = Positive Reply) employees does not obtain significance, and Mobilization obtains a significant positive association (p=.098) and Politician Mention obtains a significant negative association (p=.071). Lastly, compared to model 6, employees is significantly, positively associated with Negative Reply, and unlike in model 6 Politician Mention no longer obtains significance (p=0.45); there are no other changes in sign or significance for any of the other model variables.

Assets. In place of the number of employees, total assets was also included as a size control. In model 1 assets is negatively associated with Retweet while in model 3 assets does not obtain significance; there are no other changes in sign or significance for any of the model variables. In model 5, assets is significantly, positively associated with Negative Reply, and, compared to model 5, two control variables, Broad CSR Account Focus and URL included, fail to obtain significance; all other variables retain the same size and level of significance. In the replication of model 2, assets is negative associated (p<.01) with Retweet and one control variable, URL included, no longer obtains significance (p=0.140). Lastly, in the test of model 6 with assets included, assets does not obtain significance and three tie-building variables, User Mention, Stewardship Message, and Politician Mention, no longer obtain significance, as does the control variable Time on Twitter.

A.4.2 Alternative Dependent Variables: High Awareness.

I also ran several logits using different versions of the binary retweet dependent variable (where n=10,568). In the original variable Retweet 56% of tweets (n=10,568) receive a score of '1.' It would be useful to apply a higher threshold on the awareness generated and see whether the results hold. I thus ran additional logits with three different thresholds; specifically, at roughly the mean value of 2.08, so that the dependent variable is coded '1' if the tweet received a number of retweets at or equal to the mean value (6,302 of 18,722 tweets). I also ran the results with the threshold at 3 or more retweets (3,966 tweets) and 5 or more retweets (1,912 tweets). In all cases the sign and significance of the independent variables in all 6 models remains unchanged.

A.4.3 Alternative Regression Model: Negative Binomial Regression.

I also ran the six models using a negative binomial regression, where the dependent variables are not the binary presence of a retweet or a positive or a negative reply but rather a count of the number of, respectively, retweets and positive and negative replies. The results are substantively similar to the results presented in Table 5. Accordingly, I have chosen to concentrate on the more intuitive logit results in this study.

A.4.4 Sector: Fixed Effects and Clustered Standard Errors.

Sector Fixed Effects. It is also possible that the relationship between communication tactics and audience outcomes differs according to the industry or sector in which the company operates. The above models could, in other words, potentially suffer from omitted variable bias. The remaining sensitivity analyses serve to address such concerns. The 42 Twitter accounts represent companies working in 15 different sectors. Three of the sectors (apparel,

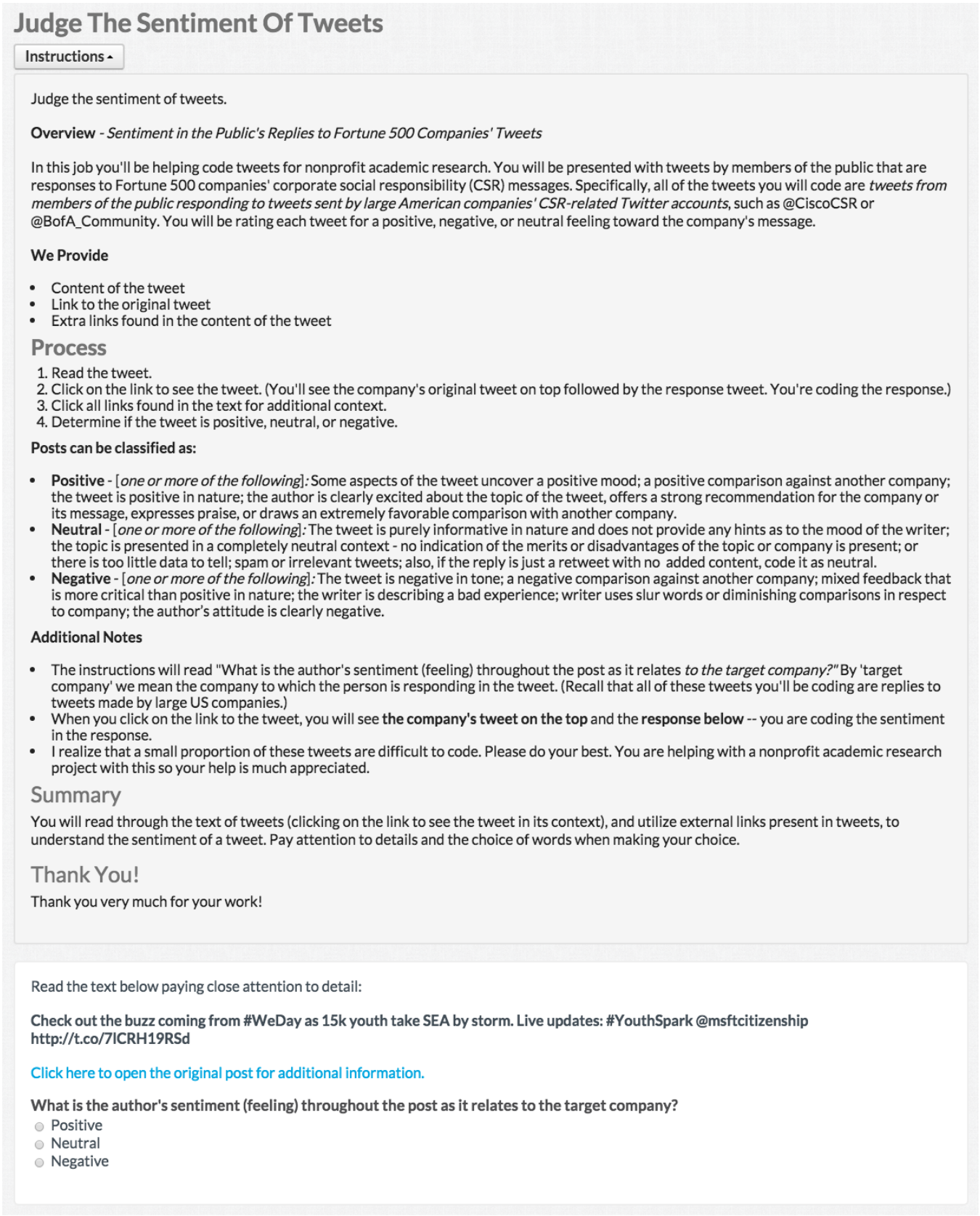
energy, and transportation) had a low number of tweets (n=18) and thus were combined into a miscellaneous category; binary variables were then created for the miscellaneous category plus each of the other 12 sectors, and a fixed effects model was run. In the replication of model 1 with industry controls, all variables are the same in terms of sign and significance save for the control variable # of Characters, which obtains signficance (p<.01). Compared to model 2, in the fixed effects regression Politician Mention gains significance (p<.01), as does Mobilization (p=.061), with no other changes in sign or significance for any model variables. In the replication of model 3, Tie-building gains significance (p=0.067) while Time on Twitter loses significance. In model 4, the only change is that Politician Mention gains significance (p= 0.083); in model 5 three controls – Time on Twitter, Broad CSR Account, and # of Characters – lose significance; while compared to model 6, Politician Mention loses significance, as do Time on Twitter, Broad CSR Account, and # of Characters.

In terms of which sector dummies were significant, the results from the replication of model 2 are illustrative. With Aerospace and Defense as the omitted category, in the replication of model 2, Miscellaneous and Financials do not obtain significance, Chemicals, Media, Motor Vehicles & Parts, Retailing, Technology, and Telecommunications are positively and signficantly related to Retweet, and Health Care, Household Products, Industrials, and Materials are negative related.

Overall, the fixed effects results are slightly more favorable to the independent variables – with minor changes to Politician Mention and Mobilization – with the sector dummies taking away some of the explained variance from the original control variables.

Standard Errors Clustered on Sector. The six models were also re-run with robust standard errors clustered on sector. Compared to the version of model 1 shown in Table 5, Time on Twitter and URL included fail to obtain significance; same with model 2. Compared to model 3, Interaction and Time on Twitter do not obtain significance, while in model Interaction, Time on Twitter and Broad CSR Account do not reach significance. Compared to model 4, Mobilization obtains a positive significant relationship (p<.01), Politician Mention obtains a negative significant relationship (p<.01), while Time on Twitter loses significance. In contrast to model 5, Interaction, Broad CSR Account and Time on Twitter no longer obtain significance. Finally, compared to model 6, Dialogue loses its significant relationship, as does Broad CSR Account and Time on Twitter. These results are slightly less favorable to the independent variables, with four instances of independent variables losing significance and two instance of independent variables gaining significance.

Sector Fixed Effects with Robust Standard Errors Clustered on Account. As a final set of robustness tests, the models were re-run with sector fixed effects and robust standard errors clustered on the Twitter account. The results are the same as the above results with standard errors clustered on sector, with the only change in sign or significance being that Mobilization and Politician Mention gain significance in model 4. Overall, the additional tests point to the robustness of the relationships between communicative tactics and the accumulation of reputational capital shown in Table 5.



Judge The Sentiment Of Tweets

Instructions

Judge the sentiment of tweets.

Overview - *Sentiment in the Public's Replies to Fortune 500 Companies' Tweets*

In this job you'll be helping code tweets for nonprofit academic research. You will be presented with tweets by members of the public that are responses to Fortune 500 companies' corporate social responsibility (CSR) messages. Specifically, all of the tweets you will code are *tweets from members of the public responding to tweets sent by large American companies' CSR-related Twitter accounts*, such as @CiscoCSR or @BofA_Community. You will be rating each tweet for a positive, negative, or neutral feeling toward the company's message.

We Provide

- Content of the tweet
- Link to the original tweet
- Extra links found in the content of the tweet

Process

1. Read the tweet.
2. Click on the link to see the tweet. (You'll see the company's original tweet on top followed by the response tweet. You're coding the response.)
3. Click all links found in the text for additional context.
4. Determine if the tweet is positive, neutral, or negative.

Posts can be classified as:

- **Positive** - [*one or more of the following*]*:* Some aspects of the tweet uncover a positive mood; a positive comparison against another company; the tweet is positive in nature; the author is clearly excited about the topic of the tweet, offers a strong recommendation for the company or its message, expresses praise, or draws an extremely favorable comparison with another company.
- **Neutral** - [*one or more of the following*]*:* The tweet is purely informative in nature and does not provide any hints as to the mood of the writer; the topic is presented in a completely neutral context - no indication of the merits or disadvantages of the topic or company is present; or there is too little data to tell; spam or irrelevant tweets; also, if the reply is just a retweet with no added content, code it as neutral.
- **Negative** - [*one or more of the following*]*:* The tweet is negative in tone; a negative comparison against another company; mixed feedback that is more critical than positive in nature; the writer is describing a bad experience; writer uses slur words or diminishing comparisons in respect to company; the author's attitude is clearly negative.

Additional Notes

- The instructions will read "What is the author's sentiment (feeling) throughout the post as it relates *to the target company?"* By 'target company' we mean the company to which the person is responding in the tweet. (Recall that all of these tweets you'll be coding are replies to tweets made by large US companies.)
- When you click on the link to the tweet, you will see **the company's tweet on the top** and the **response below** -- you are coding the sentiment in the response.
- I realize that a small proportion of these tweets are difficult to code. Please do your best. You are helping with a nonprofit academic research project with this so your help is much appreciated.

Summary

You will read through the text of tweets (clicking on the link to see the tweet in its context), and utilize external links present in tweets, to understand the sentiment of a tweet. Pay attention to details and the choice of words when making your choice.

Thank You!

Thank you very much for your work!

Read the text below paying close attention to detail:

Check out the buzz coming from #WeDay as 15k youth take SEA by storm. Live updates: #YouthSpark @msftcitizenship http://t.co/7lCRH19RSd

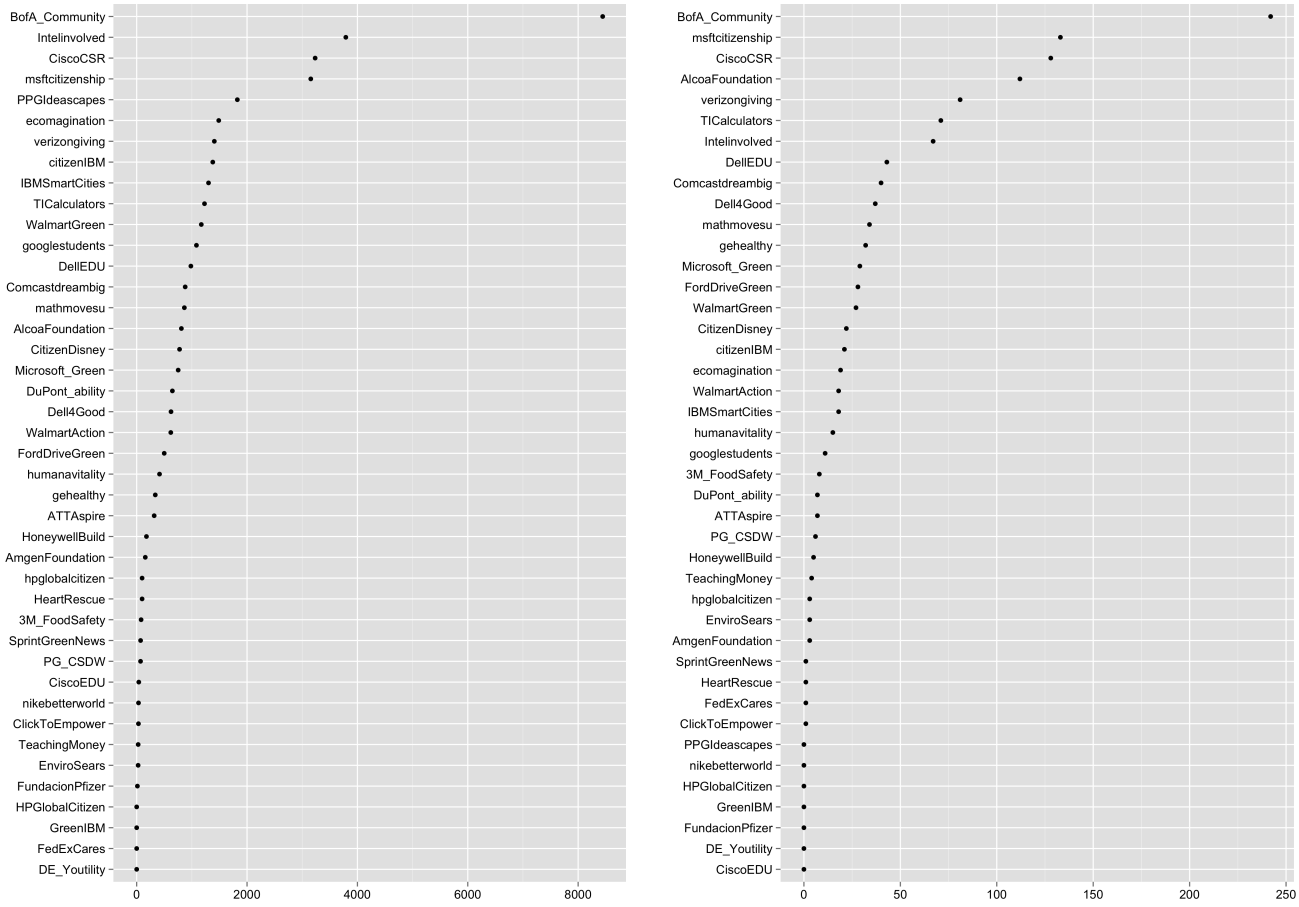
Click here to open the original post for additional information.

What is the author's sentiment (feeling) throughout the post as it relates to the target company?

- Positive
- Neutral
- Negative

Figure A.1: Instructions for Coding Sentiment Given to Crowdflower Coders

Note: Instructions with one example.



(a) Number of Retweets **(b)** Number of Positive Replies

Figure A.2: Number of Retweets and Positive Replies Received per Account in 2014

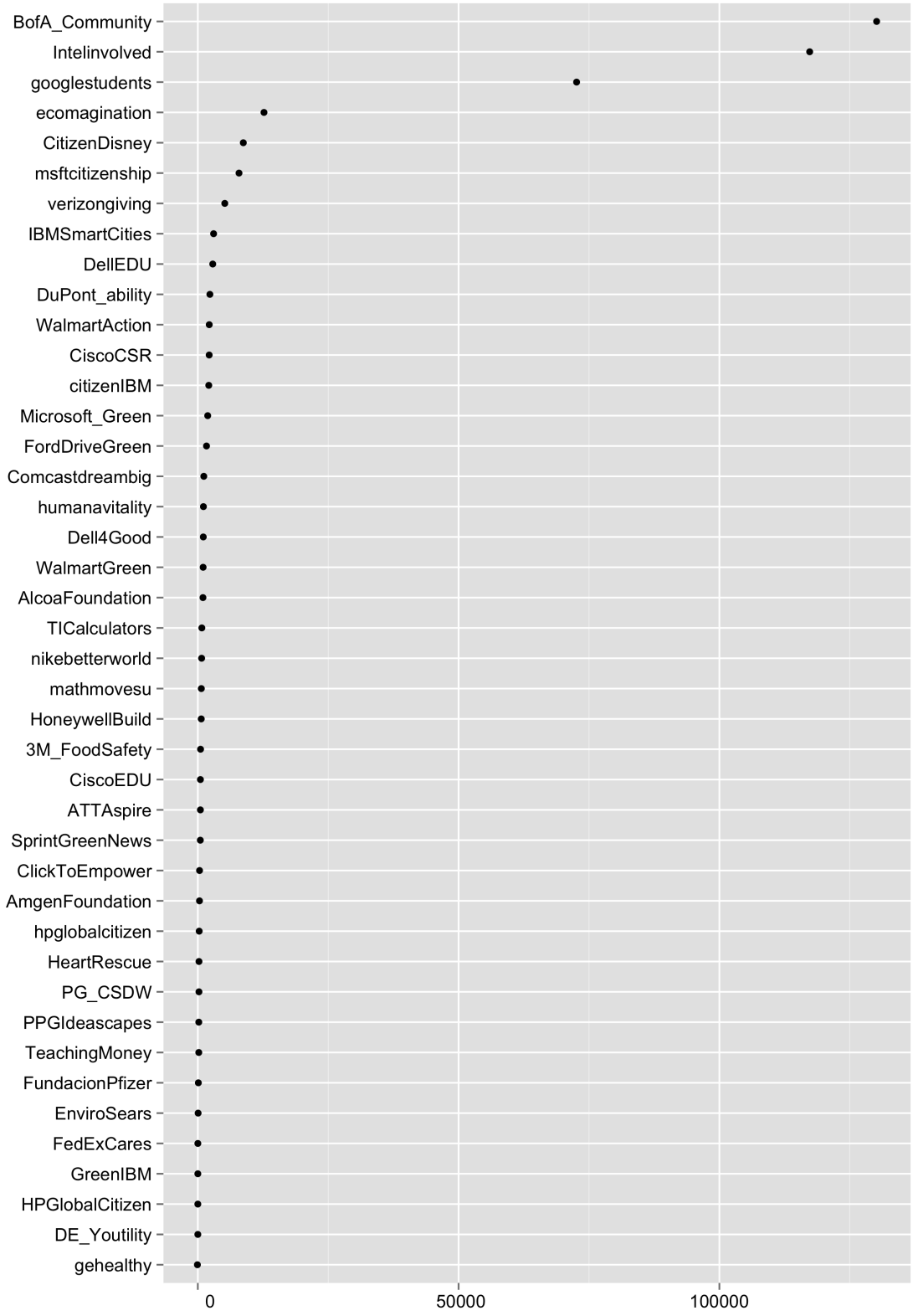


Figure A.3: Growth in Number of Followers per Twitter Account, 2014

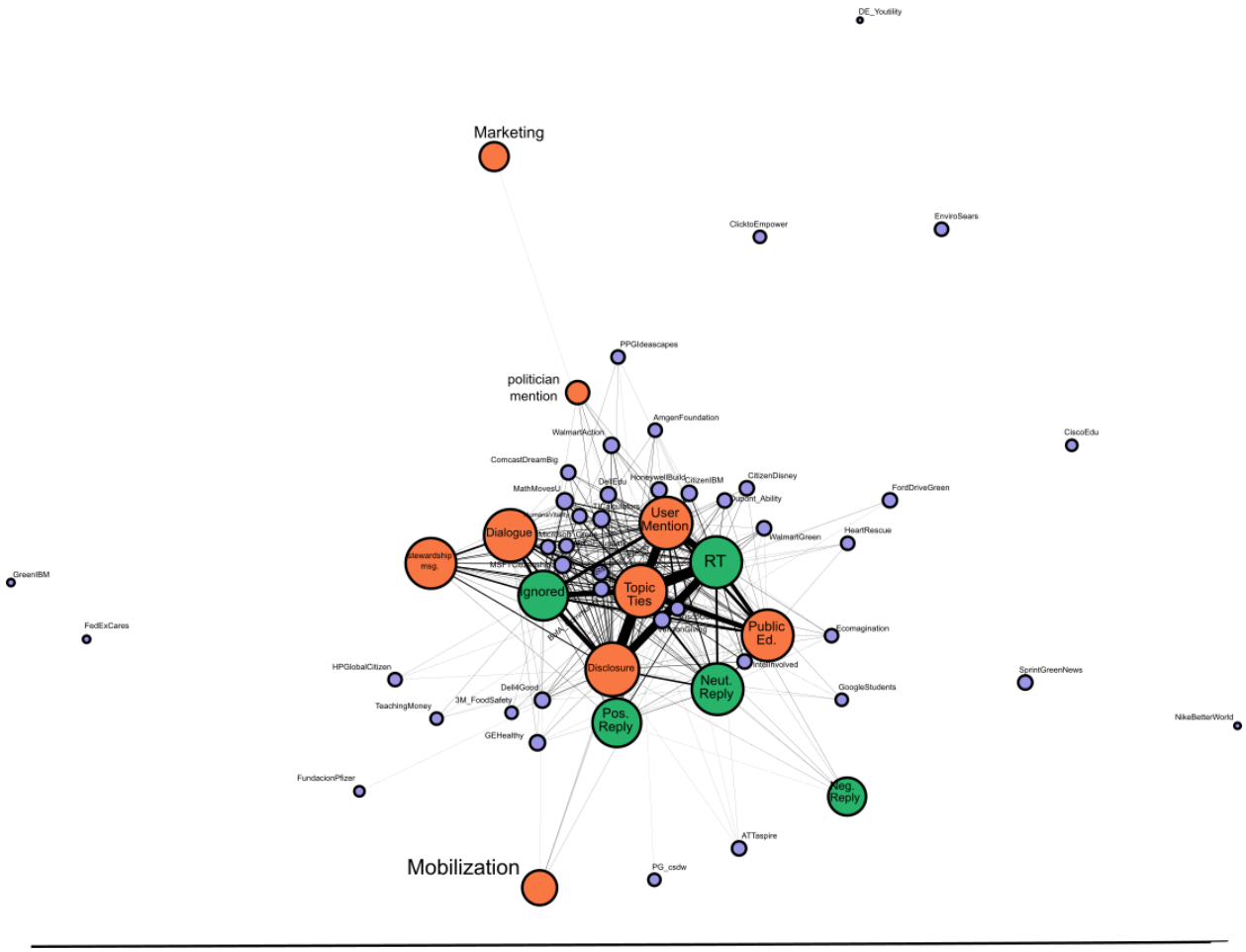


Figure A.4: Actor-Network Theory of Tactics in and Reactions to 18,722 CSR Tweets