

How Humans versus Bots React to Deceptive and Trusted News Sources: A Case Study of Active Users

Maria Glenski Tim Weninger
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana 46556
Email: {mglenski, tweninge}@nd.edu

Svitlana Volkova
Data Sciences and Analytics Group
National Security Directorate
Richland, WA 99354
Email: svitlana.volkova@pnnl.gov

Abstract—Society’s reliance on social media as a primary source of news has spawned a renewed focus on the spread of misinformation. In this work, we identify the differences in how social media accounts identified as bots react to news sources of varying credibility, regardless of the veracity of the content those sources have shared. We analyze bot and human responses annotated using a fine-grained model that labels responses as being an answer, appreciation, agreement, disagreement, an elaboration, humor, or a negative reaction. We present key findings of our analysis into the prevalence of bots, the variety and speed of bot and human reactions, and the disparity in authorship of reaction tweets between these two sub-populations. We observe that bots are responsible for 9-15% of the reactions to sources of any given type but comprise only 7-10% of accounts responsible for reaction-tweets; trusted news sources have the highest proportion of humans who reacted; bots respond with significantly shorter delays than humans when posting answer-reactions in response to sources identified as propaganda. Finally, we report significantly different inequality levels in reaction rates for accounts identified as bots vs not.

I. INTRODUCTION

Misinformation spread in social networks has become a critical focus as users rely on these platforms as a primary source of news [11]. Current studies in this area have focused on rumor and misinformation detection with a primary focus on the network’s role in information diffusion models [9], [10], [13], [26]. Other studies compare the behavior of traditional and alternative media [17], classify media sources into sub-categories of misinformation [22], or attempt to detect rumor-spreading users [14]. These and other studies have found that the size and shape of (mis)information cascades within a social network depends heavily on the initial reactions of the users. Yet, we still lack an understanding of how users (human and automated alike) react to news sources of varying credibility and how their various response types contribute to the spread of (mis)information. The present work aims to fill this gap by labelling bot and human users’ reactions to (mis)information posted by various news sources to measure how bot and human user reactions to deceptive news sources differ from their responses to trusted news sources.

Instead of focusing on user reactions to individual news stories, the current work compares human-user and bot reactions to news sources of varying credibility. We focus on how behavior of bot and human users differ in four specific areas:

1) concentration of reactions to news sources of each level of credibility, *i.e.*, are bots responsible for a larger proportion of the reactions for one class of news sources over another? (*prevalence of bots*), 2) the variety of reactions each class of news sources evoke, (*reaction variety*), 3) the speed with which reactions are posted, (*reaction speed*), and 4) how equally the volume of reactions are spread across the set of users who reacted (*reaction inequality*).

II. RELATED WORK

Prevalence of Bots. Previous studies have identified the widespread presence of automated accounts or “bots” on social media. A 2014 filing from Twitter acknowledged that 8.5 percent of its active monthly users were automated accounts¹ and subsequent studies found this to be a low estimate of the actual prevalence of bot accounts [1], [20]. Recent work has found that accounts spreading disinformation are significantly more likely to be automated accounts [16]. Other studies highlight evidence of bot participation in political discussion [5], [12], [25] and astroturf campaigns that present the appearance of widespread support of a candidate, opinion, or topic artificially [15]. A 2018 Pew Research center study found that the majority (66%) of links tweeted to popular news sites are posted by accounts that are likely to be bots, *i.e.*, whose behavior is more similar to bot accounts than to humans [24]. We seek to answer whether similar trends hold among reactions to news sources.

Reaction Variety. Linguistic markers have been found to be effective for early detection of rumors in social networks. For example, Kwon et al. [9] demonstrated better detection performance of rumors on Twitter by using user and linguistic features rather than structural or temporal network features. Similarly, Zhao et al. [29] identified clusters of tweets that contain disputed claims by searching for fact-checking language. Recently, Zhang et al. [28] classified Reddit comments into eight types including agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, and question, and analyzed patterns from these discussions arranged by

¹https://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm?_ga=1.155500795.1900968760.1407851022

various subreddits. Our work goes one step further and employs information credibility classifiers like those mentioned above in order to better understand how (and how fast) human users and bots react to information posted by news sources of varying credibility.

Reaction Speed. Information diffusion studies have often used epidemiological models, originally formulated to model the spread of disease within a population, in the context of social media [6], [18], [27]. In this context, users are *infected* when they spread information to other users. A recent study by Vosoughi et al. [23] found that news that was fact-checked (post-hoc) and found to be false had spread faster and to more people than news items that were fact-checked and found to be true. In this work, we examine the speed at which users react to content posted by news sources of varying credibility and compare the delays of different types of responses. By contrasting the speed of reactions of different types, from different types of users (bot and human), and in response to sources of varying credibility, we are able to determine whether deceptive or trusted *sources* have slower immediate share-times overall and within each combination of user, reaction, and news source types.

Reaction Inequality. In the context of social media, the 1% rule and its variants indicate that most users only browse content while a mere 1% of users contribute new content [4], [19]. Within the subset of those who actively contribute new content, Kumar and Geethakumari [8] found a larger disparity among users who retweeted news from sources that were identified as spreading disinformation. That is, a small number of highly active users were responsible for the vast majority of retweets of disinformation. This study focused only on keywords related to the events in Egypt and Syria in 2013. To answer this research question more generally, the present work quantifies and compares the disparity in sharing behavior of users who frequently reacted to news sources across the various categories of sources, in particular the disparity within each of the reaction types. Specifically, for each type of reaction and each type of news source, we examine whether reactions from bots and human users who frequently reacted are equally distributed across the population of users or if there are a small group of vocal users responsible for the majority of the reaction-tweets.

III. DATA COLLECTION AND ANNOTATION

Deceptive news sources that primarily share clickbait, conspiracy theories, or propaganda were previously collected by Volkova et al. [22] from several public resources that annotate suspicious news accounts.² The authors also compiled a set of trusted news sources that tweet in English with Twitter-verified accounts which were manually labeled. We collected a set of news sources from <https://euvsdisinfo.eu/> that were identified as a source of disinformation by the European Union’s East Strategic Communications Task Force. As of

November 2016, EUvsDisinfo reports include almost 1,992 confirmed disinformation campaigns found in news reports from around Europe and beyond. We limited our set to news sources identified between 2015 and 2016 [21].

In total, we focused on 282 news sources which were identified as sources who spread:

- **trusted news (T):** factual information with no intent to deceive the audience;
- **clickbait (CB):** attention-grabbing, misleading, or vague headlines to attract an audience;
- **conspiracy theories (CS):** uncorroborated or unreliable information to explain events or circumstances;
- **propaganda (P):** intentionally misleading information to advance a social or political agenda; or
- **disinformation (D):** fabricated and factually incorrect information meant to intentionally deceive the audience.

We collected tweets posted between January 2016 and January 2017 that explicitly @mentioned or directly retweeted content from one of our 282 sources via the public Twitter API and assigned a label to each tweet based on the class of the source @mentioned or retweeted. Then, we focused on the subset of 4,613,517 tweets identified as English-content in the Twitter metadata. We further focused on users who frequently interacted (at least five times) with the news sources we considered, using tweets posted in any language, which resulted in 431,771 English-tweets for 255 news sources from 184,248 distinct, frequently interacting users. We then classified each of the reaction-tweets as an agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, question, or other. To do so, we used linguistically-infused neural network models [3] trained on a manually annotated reaction dataset from Zhang et al. [28].

Finally, we gathered botometer scores [2] for each user who posted a reaction-tweet and partitioned the data into bot reactions and human-user reactions using a bot-score threshold of 0.5. That is, human-user reactions were posted by users with a bot score under the threshold of 0.5 and the bot reactions dataset comprises tweets posted by users with bot scores at or above the threshold. A summary of the dataset across source types is presented in Table I.

TABLE I
SUMMARY OF ENGLISH REACTIONS FROM USERS WHO REACTED FREQUENTLY (≥ 5 REACTIONS BETWEEN JAN 2016 AND JAN 2017).

Source-Type	Sources		Reactions	
	# Accounts	# Tweets	# Users	# Tweets
Trusted	173	1,633,996	173,098	2,875,120
Clickbait	10	13,764	8,088	22,352
Conspiracy	13	31,584	14,047	80,025
Propaganda	25	81,305	51,160	295,070
Disinformation	34	68,319	26,131	164,040

IV. METHODOLOGY

In this section, we describe the methodology we used to examine the behavior of bot and human user accounts across

²Deceptive news lists include <http://www.fakenewswatch.com/>, <http://www.propornot.com/p/the-list.html>.

varying reactions and in reaction to news sources of varying levels of credibility. As previously discussed, we focus on four specific types of behavior: prevalence of bots, reaction variety, reaction speed, and the inequality of reaction volume.

First, we examine the prevalence of bots, *i.e.*, the relative presence of bots in reactions to news sources of each type. We consider the following two distributions: 1) bot scores of users who reacted to news sources of a given type and 2) bot scores associated with reaction-tweets (the bot scores of users who posted the reaction). The distribution of reaction-users focuses on the distribution of bot scores over the set of unique users who reacted, each user is represented once and only once. On the other hand, users may be represented multiple times in the distribution of bot scores associated with reaction-tweets, if they reacted to a news source of a given class multiple times. With these two distributions of bot scores, we are able to examine the prevalence of bots within the population of reacting users and within the population of reactions broadcast.

As a result of our bot classification methodology, we are able to examine user types using coarse and fine-grained classifications. We first examine the distributions of bots and humans users at a coarse granularity with a binary classification of users as either a bot or human user account. Then, we consider a fine-grained distinction using the bot scores of users and compare the distributions of bot scores for users who react and of bot scores associated with reaction-tweets (*i.e.*, the bot score of the user who posted). Mann Whitney U (MWU) tests that compare distributions across types of sources and types of users are used to identify statistically significant differences in these fine-grained distributions.

The next characteristic that we evaluate is the variety of reactions each class of news source elicits from bots and from human users. We compare distributions across reaction types overall and separated them into each category of user. Comparisons of reaction variety within each user type allows us to identify certain reactions, classes of news sources, or reactions to a class of news source that have higher concentrations of bot (or human) reactions. Then we consider the tendency of each user type by comparing the frequencies of each reaction type across all classes of sources between bot and human users.

Next we examine the speed of reactions. To answer whether how quickly bots or human users react differs or whether users react to content from trusted sources faster than from deceptive sources, we look at reaction delays for each user type, reaction type, and response to each class of news sources. We define the reaction delay as the time elapsed between the source tweet and when the reaction occurred. We compare the cumulative distribution functions (CDFs) of each user type within and across each type of source to analyze the delay patterns.

Finally, we compare the inequality in reactions among bots and human users. That is, how evenly the volume of reaction-tweets is spread across users of each type; Does each user post an equal number of reactions? We do so using two measures that have been commonly used to measure income inequality: Lorenz curves and Gini coefficients. Rather than measure how much of the total population's income each individual

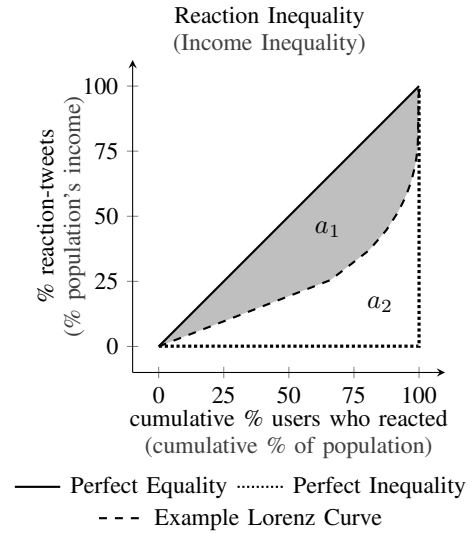


Fig. 1. Lorenz curves and Gini coefficients. As a graphical representation of income inequality within a population, Lorenz curves plot the share of income by the cumulative share of the population. The Gini coefficient is the proportion of the area under the line of perfect equality ($a_1 + a_2$) that is captured between the line of perfect equality and the Lorenz curve (a_1). We adapt Lorenz curves to measure the inequality in reaction volume by plotting the share of the total reaction volume, *i.e.*, the y % of reaction-tweets posted, by the share of the population who reacted, *i.e.*, the cumulative x % of users ordered by least to most reaction-tweets posted.

is responsible for, we repurpose these metrics to measure how much of the total reaction-tweet volume each user is responsible for. This allows us to compare *reaction inequality* across source types the way that economists compare income inequality across countries or regions.

Lorenz curves have traditionally been used to illustrate the distribution of income or wealth graphically [7]. In those domains, the curves plot the cumulative percentage of wealth or income compared against the cumulative (in increasing shares) percentage of a corresponding population. The degree to which a Lorenz curve deviates from the straight diagonal line ($y = x$) representative of perfect equality represents the inequality present in the distribution. In our case, the Lorenz curve is adapted to illustrate the cumulative percentage of propagation (tweets shared) as a function of the cumulative percentage of users posting, as shown in Figure 1.

$$\hat{G} = 1 - \sum_{k=1}^n (X_k - X_{k-1})(Y_k + Y_{k-1}) \quad (1)$$

The Gini coefficient is defined as the proportion of the area under the line of perfect equality that is captured above the Lorenz curve, *i.e.*, $\frac{a_1}{a_1 + a_2}$ in Figure 1. The Gini coefficients reported in subsequent sections are calculated using the formula in Eq. 1, which is an approximation of the points of the Lorenz curves observed in the collected data. Using income as an example, Gini coefficients can grow larger than 1 but only if individuals within the population can be responsible for negative shares, that is, if individuals can have negative incomes. In our data, users must be responsible for at least 1

TABLE II

(PREVALENCE OF BOTS) DISTRIBUTIONS ACROSS BOT ACCOUNTS (BOT SCORE ≥ 0.5), HUMAN ACCOUNTS (BOT SCORE < 0.5), AND UNKNOWN ACCOUNTS (FOR WHICH WE COULD NOT COLLECT A BOT SCORE) WITHIN THE SET OF USERS WHO REACTED (U) AND THE SET OF REACTION TWEETS (T) FOR EACH CLASS OF NEWS SOURCE. HIGHEST PROPORTIONS OF EACH USER TYPE ARE HIGHLIGHTED IN BOLD AND LOWEST PROPORTIONS ARE IN ITALICS.

Source-Type	% Bot		% Human		% Unknown	
	U	T	U	T	U	T
Trusted	7.47	12.57	77.32	74.46	<i>15.22</i>	<i>13.03</i>
Clickbait	10.17	15.06	74.10	72.62	15.73	<i>12.35</i>
Conspiracy	7.90	<i>8.90</i>	72.79	76.50	19.31	14.62
Propaganda	6.80	11.54	75.00	70.56	18.20	17.94
Disinformation	9.64	13.29	73.11	<i>70.18</i>	17.25	16.65

reaction-tweet in order to be considered part of the dataset, so Gini coefficients in our analysis have an upper-bound of 1.

V. ANALYSIS

Here we present the key results of our analysis of the behavior of bots and human users in reaction to news sources of varying credibility: the prevalence of reactions from bots and the variety, speed, and the inequality in volume of reaction tweets evoked by each class of news source.

A. Prevalence of Bots

In this subsection, we consider the prevalence of bot users among the audience and reactions broadcast to the community. The distributions of users across bot, human, and unknown (accounts for which we could not collect bot scores) within each class of news source are presented in Table II.

As shown in Table II, bots are responsible for approximately 9-15% of the reactions to sources of any given type but only comprise around 7-10% of users responsible for reaction-tweets. We see that although conspiracy sources have the lowest presence of human users within the population of users who react, they have the highest proportion of reactions authored by human-users. *Trusted news sources have the highest relative presence of human users.* Interestingly, disinformation news sources have only the second highest proportions of bots for users who reacted as well as reaction tweets posted. Instead, clickbait news sources have the highest presence of bots with 10.17% of users who were responsible for 15.06% of the reaction-tweets for clickbait sources identified as bots.

Figure 2 illustrates the distributions of bot scores of users who reacted (left) and the scores associated with reaction-tweets, *i.e.*, the bot score of the user who posted the tweet, (right). When we compare distributions of users' bot scores across classes of news sources, we find statistically significant differences. Mann Whitney U comparisons identified significant ($p < 0.01$) differences between distributions for clickbait and trusted or propaganda news sources, where reactions and users who post reactions to clickbait sources have higher bot scores, on average, than trusted or propaganda news sources. Although the distributions of bot scores of unique users and scores associated with reaction tweets are not statistically

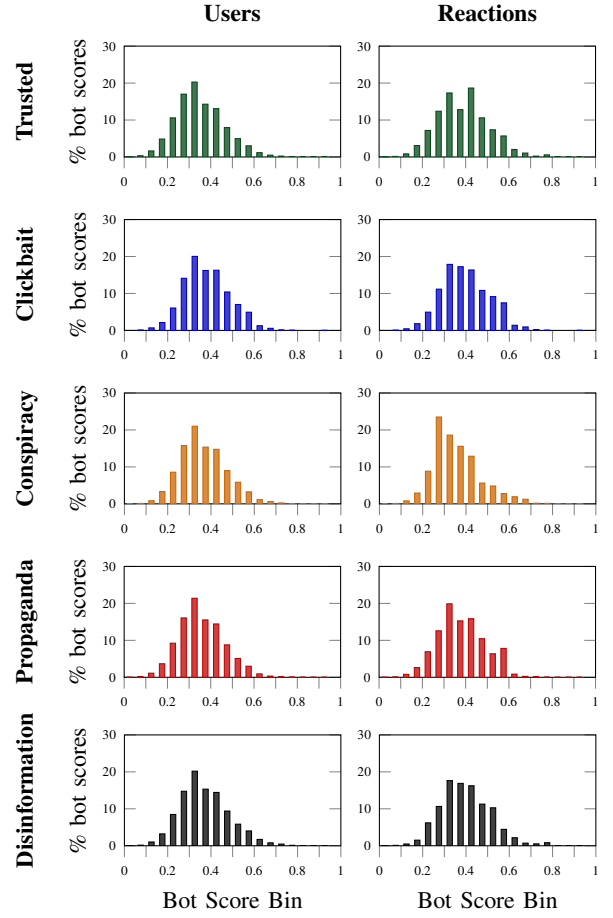


Fig. 2. (Prevalence of Bots) Bot score distributions, using a bin width of 0.05, for users who reacted (left) and reaction-tweets (right). Mann Whitney U comparisons of raw distributions found that the average bot score of a user who posted a reaction-tweet is higher ($p < 0.01$) than the average bot score of a user who reacted for all source types except for Conspiracy-sources, where the average bot score of a user who posted a reaction-tweet is lower ($p < 0.01$).

significant, the slight changes in the shape of the distributions, *e.g.*, between the two distributions for Conspiracy sources, paired with the discrepancies in Table II hint at the inequality of reaction tweet volume. That is, they indicate that reactions are not evenly spread across users. We investigate this further in our analysis of reaction inequality.

B. Reaction Variety

We plot the distributions of reaction-types for each of the five classes of news sources in Figure 3 and the distribution across bot, human, and unknown users for each source class and reaction type combination for the most frequent reaction types in Table III. When we compare the distributions of reaction types, we see that the most common reaction types (*i.e.*, present in $\geq 10\%$ of reactions) are answer, elaboration, question, and “other” across all classes of media. In Figure 4 we present the relative frequencies of the most common reactions within the reaction-tweets posted by a given user type in response to news sources of a given class. These plots

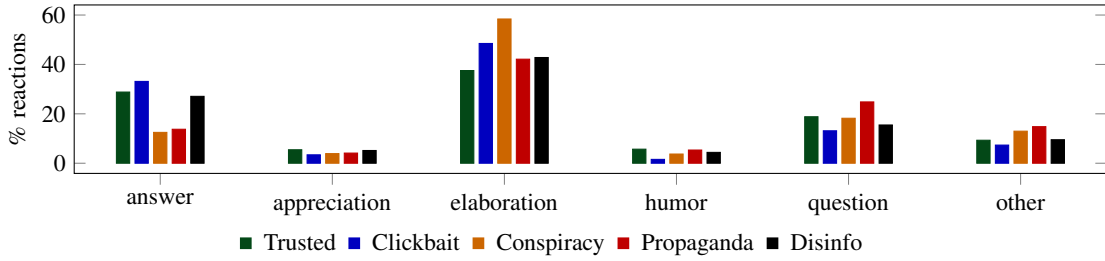


Fig. 3. (Reaction Variety) Distributions of predicted reaction-types within tweets that directly responded to sources of each source-type.

TABLE III
(REACTION VARIETY) PROPORTIONS OF REACTIONS POSTED BY BOT, HUMAN, OR UNKNOWN USERS FOR EACH SOURCE CLASS AND REACTION TYPE COMBINATION FOR THE MOST FREQUENT REACTION TYPES. SOURCE CLASS(ES) WITH THE LOWEST PROPORTIONS FOR EACH USER TYPE ARE HIGHLIGHTED WITH BOLD FOR EACH OF THE REACTION TYPES.

	Answer			Elaboration			Question			Other		
	B	H	U	B	H	U	B	H	U	B	H	U
Trusted	0.16	0.69	0.15	0.10	0.77	0.13	0.12	0.75	0.13	0.12	0.74	0.14
Clickbait	0.24	0.66	0.11	0.11	0.76	0.13	0.12	0.75	0.13	0.12	0.76	0.12
Conspiracy	0.09	0.75	0.16	0.09	0.77	0.14	0.08	0.79	0.13	0.09	0.74	0.17
Propaganda	0.25	0.57	0.17	0.09	0.73	0.18	0.10	0.71	0.19	0.09	0.73	0.19
Disinformation	0.15	0.68	0.17	0.12	0.71	0.16	0.12	0.71	0.16	0.14	0.70	0.16

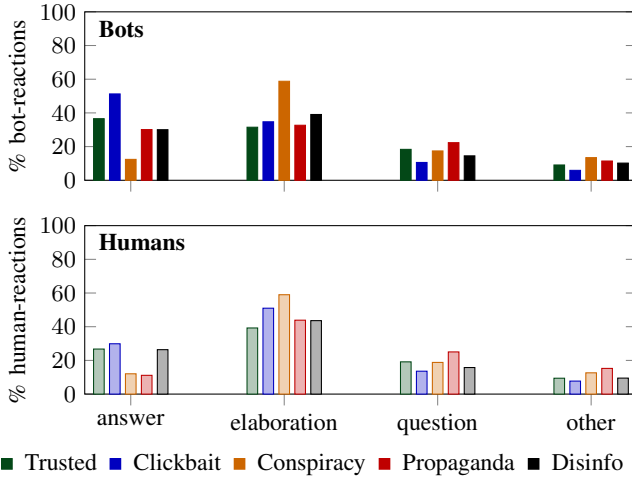


Fig. 4. (Reaction Variety) Frequencies of most common reaction-types within reactions to news sources of each class posted by bot accounts (above) and human user accounts (below), as a percentage of reactions posted by accounts within each population.

focus more closely on how reaction frequencies differ within a single user-type population.

When we examine the distributions of each class, we find several key differences in the variety of reactions elicited. Conspiracy news sources have the highest relative rate of elaboration responses, *i.e.*, “*On the next day, radiation level has gone up. [url]*”, with a more pronounced difference within the bot population. Conspiracy news sources also have the lowest relative rate of answer reactions within the bot population, but not within human users. Clickbait news sources, on the other hand, have the highest relative rate of answer reactions and the lowest rate of question reactions across both populations of user types.

Conspiracy and propaganda news sources have higher rates of human question-reactions than they do human answer-reactions; human users who react to these types of news sources question content from the source more often than they respond with an answer. While we see a similar trend within human users for conspiracy sources, we see a higher relative rate of answer reactions to propaganda sources when we examine relative rates of bot reactions.

C. Reaction Speed

Next, we study the speed with which bot and human users react to news sources. CDF plots for reaction delays of the most frequently occurring reactions are shown in Figure 5. These plots illustrate the percentage of reactions that occur within the first x hours after a source posted the original content users reacted to. As expected, a large proportion of the reaction activity occurs soon after a news source posts across all reaction and source type combinations.

Mann Whitney U tests that compared distributions of reaction delays found that humans elaborate on and question content from clickbait sources faster than bots do ($p < 0.01$). This is reflected in Figure 5 where we see the CDF curve for humans pulls above the curve for bots due to the heavier concentration (at least 80%) of reactions with very short (≤ 6 hours) delays, compared to bot users with approximately 60-70% of reactions that occurred within the first 6 hours. We see similar trends for all the other combinations of reaction and source types but a few notable exceptions. *In the case of answer-reactions in response to content from propaganda news sources, bots respond with significantly shorter delays than human users do* ($p < 0.01$). MWU tests comparing bot and human answer-reactions to clickbait and disinformation sources were not found to differ with statistical significance.

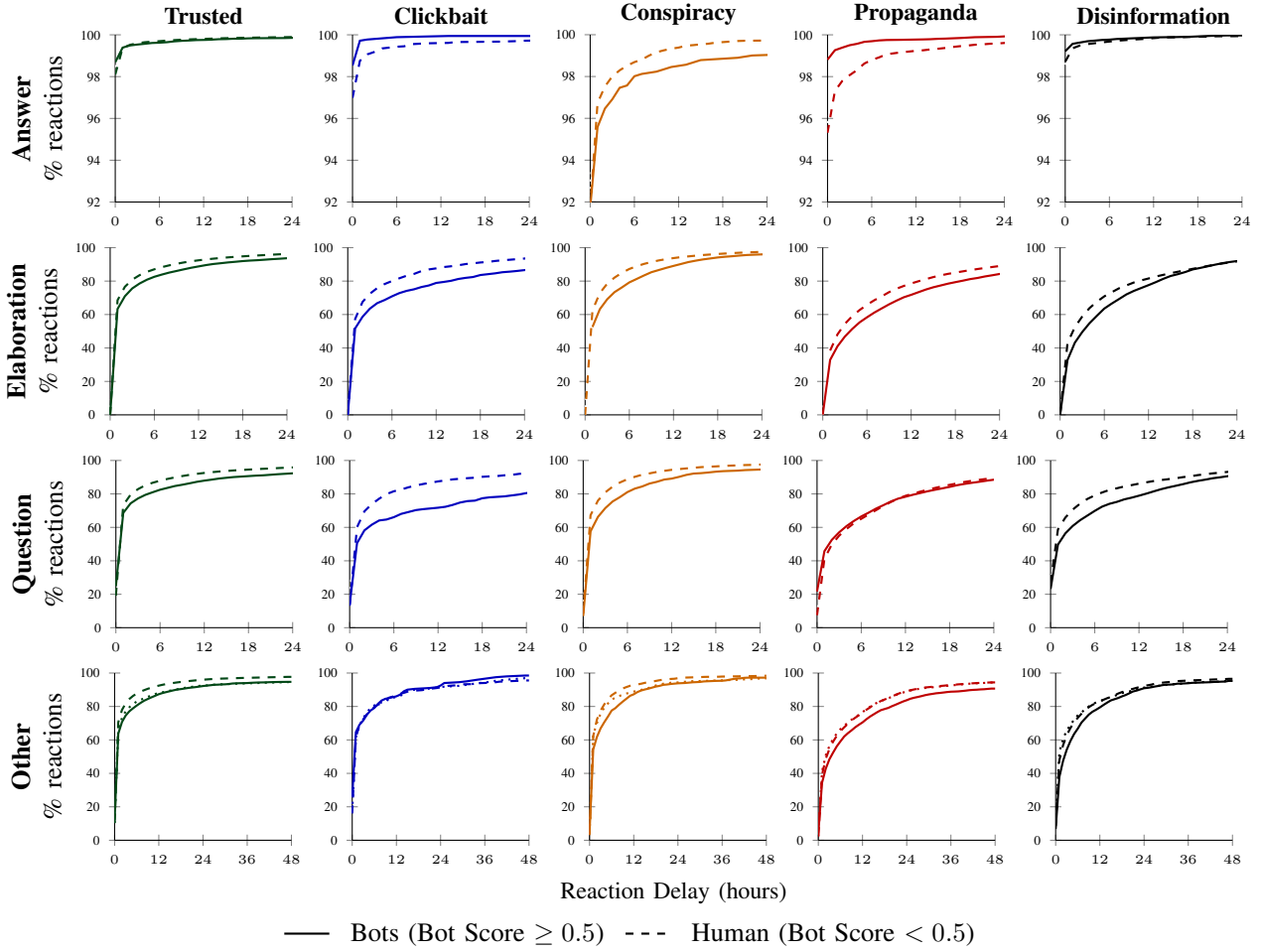


Fig. 5. (Reaction Speed) Cumulative distribution function (CDF) plots of the volumes of reactions by reaction delays in hours (*i.e.*, the delay between when a source posted content and when the reaction tweet was posted) for bots and human user accounts for the most frequently occurring reactions (occurring in at least 10% of tweets) for each source-type, using a step size of one day.

D. Reaction Inequality

Finally, we investigate reaction inequality to answer the question: does each user share an equal number of reactions, or are some user or users responsible for a disproportionate number of the reaction tweets for each of the most common reaction types (answer, elaboration, question, and “other”) ? In Figure 6 we present the Lorenz curves for bots and human users when we consider populations with reaction tweets for each combination of reaction type and class of source.

There are significant differences (MWU $p < 0.01$) between the Lorenz curves for bot and human users for all combinations of reaction and source types except for elaboration reactions to clickbait news sources and elaboration, question, and “other” reactions to conspiracy sources. In these cases, human users are also unevenly responsible for reaction tweets, *i.e.*, a subset of the human users are responsible for a disproportionate number of the human-reactions, and the disparity between users who react infrequently and those who post a substantial number of reactions is similar to those within the corresponding populations of bot users.

When users reacted to conspiracy sources, the volume

of reaction tweets are similarly unequally distributed across users within the populations of bots and human users except for answer-reactions. Answer-reactions posted in response to conspiracy sources have a smaller prolific subset of bot users responsible for an unexpectedly large volume of the reaction tweets. Human users also respond unevenly with a subset of users who post a disproportionate amount of the reactions, but to a lesser extent than the population of bot users who posted reactions. We see similar patterns across all significant comparisons. That is, *bot populations, if significantly different from the corresponding human user population, always have a higher level of disparity in reaction volumes than the corresponding human users.*

Table IV presents the increases in Gini coefficient from the human user to bot populations. For clarity, we present only the significant increases ($p < 0.05$) with dashes (—) in place of results without significance. Increases are presented in both absolute terms and relative to the Gini coefficient of the human user population. The most extreme difference is seen in answer-reaction to propaganda sources, with the bot population having a Gini coefficient 58.6% (+0.34) larger than

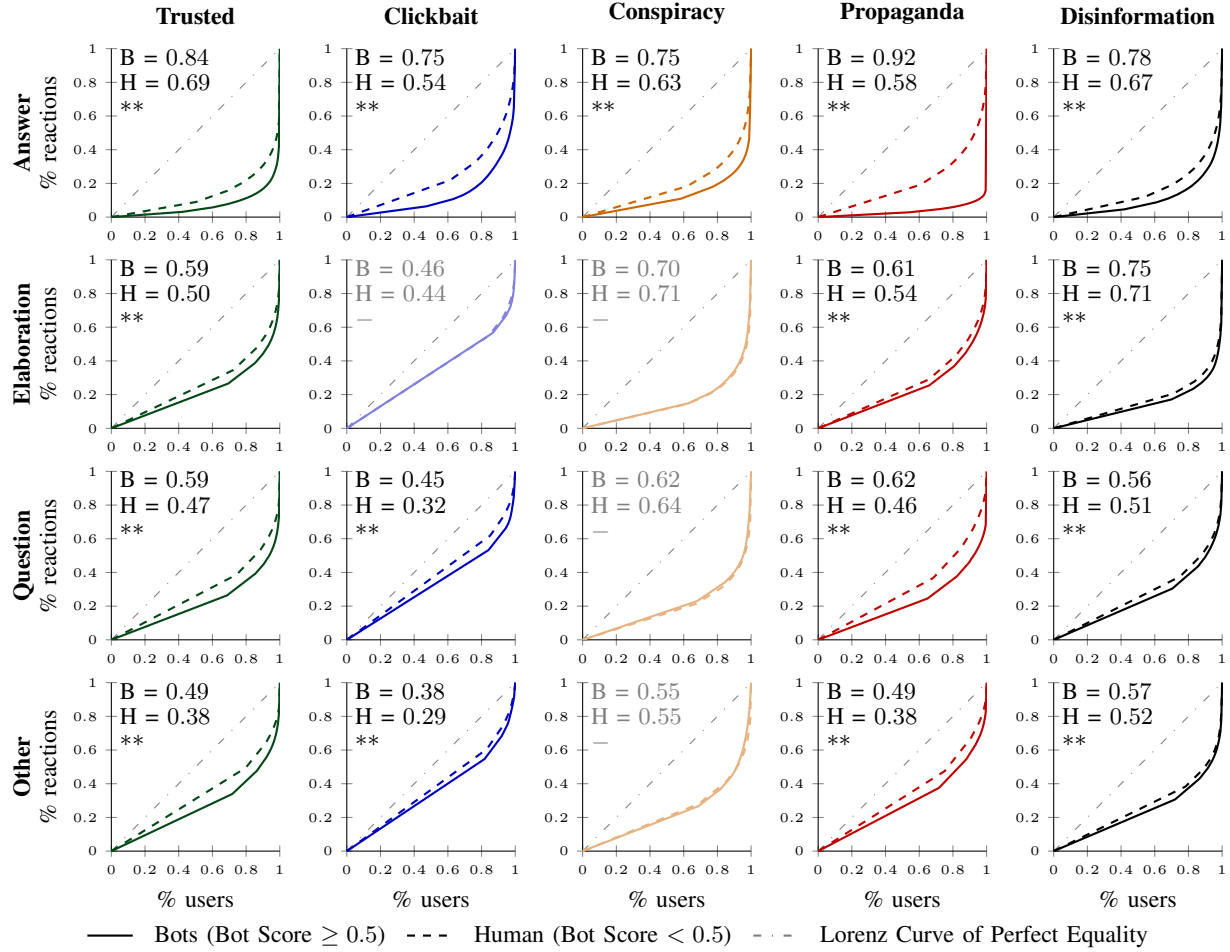


Fig. 6. (Reaction Inequality) Lorenz curves for each of the frequently occurring reactions (occurring in at least 5% of tweets) for each source-type. These Lorenz curves plot the share of reactions by the cumulative share of the population (bots, humans, or accounts without bot scores) as a graphical representation of inequality in reaction volume within each population. The gray dash-dotted line reflects the Lorenz curve that would result from a population wherein each user was responsible for an equal number of reactions. Gini coefficients for bot (B) and human (H) accounts and statistical significance results of Mann Whitney U (MWU) comparisons of Lorenz curves are listed in the top left corner of each subplot. ** if $p < 0.01$, * if $p < 0.05$, and — if $p \geq 0.05$. Lorenz curves and Gini coefficients are presented faded where there are no significant differences between bot and human users.

TABLE IV

(REACTION INEQUALITY) THE DIFFERENCE (Δ) IF STATISTICALLY SIGNIFICANT (MWU $p < 0.01$) BETWEEN GINI COEFFICIENTS FOR BOT (B) AND HUMAN (H) USER ACCOUNTS AND THE RELATIVE INCREASE ($\% \Delta$) FROM THE HUMAN USER TO BOT GINI COEFFICIENT, *i.e.*, $(B-H)/H$. A DASH (—) IS SHOWN IF NO SIGNIFICANT DIFFERENCE WAS FOUND ($p \geq 0.05$). HIGHEST RELATIVE INCREASES ARE HIGHLIGHTED IN BOLD WITHIN SOURCE TYPES AND ITALIZED WITHIN REACTION TYPES.

Reaction	Trusted		Clickbait		Conspiracy		Propaganda		Disinfo	
	Δ	$\% \Delta$	Δ	$\% \Delta$	Δ	$\% \Delta$	Δ	$\% \Delta$	Δ	$\% \Delta$
Answer	0.15	21.74	0.21	38.89	0.12	19.05	0.34	58.62	0.11	16.42
Elaboration	0.09	<i>18.00</i>	—	—	—	—	0.07	12.96	0.04	5.63
Question	0.12	25.53	0.13	40.63	—	—	0.16	34.78	0.05	9.80
Other	0.11	28.95	0.09	<i>31.03</i>	—	—	0.10	25.64	0.05	9.62

human users do. We find that the highest relative increases for the more deceptive news source classes (conspiracy, propaganda, and disinformation) occur when we compared answer-reactions. The highest relative increase for elaboration reactions occurs within elaboration-reactions to trusted news sources. The highest relative increase in inequality for reactions to trusted news source, however, occurs within the class of “other” reactions, *i.e.*, reactions that our annotation model

did not predict to be one of the eight reaction types. In contrast, we see the lowest significant relative differences between human and bot users in reactions to disinformation sources. We see that the Gini coefficients for bots are only 5.6% higher than humans for elaboration-reactions, and approximately 10% higher for both question-reactions and other-reactions.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a novel analysis of bot and human-user reactions to sources of varying levels of credibility using fine-grained reaction labels. We identified several key differences in the prevalence of bots within reactions and populations of users who reacted, the variety of reactions each news source evokes, the speed with which different reactions occurred and the inequality of participation in the set of reactions. Future work will focus on further exploration of the differences in evolution of the response to deceptive sources, an expanded analysis that incorporates both frequent and infrequently reacting users, and comparisons across multiple platforms *e.g.*, Facebook and Reddit.

ACKNOWLEDGMENTS

Twitter data used for the analysis in this paper was collected using public Twitter API and analyzed over the period of 01/2016 – 01/2017. Botometer data was collected by the University of Notre Dame using public APIs. The research was supported by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. This research is also supported by the Defense Advanced Research Projects Agency (DARPA), contract W911NF-17-C-0094. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

- [1] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: Human, bot, or cyborg?" in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. ACM, 2010, pp. 21–30.
- [2] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 273–274.
- [3] M. Glenski, T. Weninger, and S. Volkova, "Identifying and understanding user reactions to deceptive and trusted social news sources," in *ACL*, 2018.
- [4] E. Hargittai and G. Walejko, "The participation divide: content creation and sharing in the digital age," *Information, Community and Society*, vol. 11, no. 2, pp. 239–256, 2008.
- [5] P. N. Howard and B. Kollany, "Bots, #strongerin and #brexit: Computational propaganda during the uk-eu referendum," *Social Science Research Network*, 2016.
- [6] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Social Network Mining and Analysis*, 2013.
- [7] N. C. Kakwani and N. Podder, "On the estimation of lorenz curves from grouped observations," *International Economic Review*, pp. 278–292, 1973.
- [8] K. K. Kumar and G. Geethakumari, "Detecting misinformation in online social networks using cognitive psychology," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, p. 14, 2014.
- [9] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PLoS one*, vol. 12, no. 1, p. e0168344, 2017.
- [10] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proceedings of ICDM*, 2013.
- [11] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [12] P. T. Metaxas and E. Mustafaraj, "Social media and the elections," *Science*, vol. 338, no. 6106, pp. 472–473, 2012.
- [13] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of EMNLP*, 2011.
- [14] B. Rath, W. Gao, J. Ma, and J. Srivastava, "From retweet to believability: Utilizing trust to identify rumor spreaders on twitter," in *Proceedings of ASONAM*, 2017.
- [15] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," *ICWSM*, vol. 11, pp. 297–304, 2011.
- [16] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of fake news by social bots," *arXiv preprint arXiv:1707.07592*, 2017.
- [17] K. Starbird, "Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter," in *ICWSM*, 2017.
- [18] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer, "Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks," in *WWW*, 2015.
- [19] T. van Mierlo, "The 1% rule in four digital health social networks: An observational study," *J Med Internet Res*, vol. 16, no. 2, p. e33, Feb 2014.
- [20] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *ICWSM*, 2017.
- [21] S. Volkova and J. Y. Jang, "Misleading or falsification: Inferring deceptive strategies and types in online news and social media," in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 575–583.
- [22] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," in *ACL*, 2017.
- [23] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. [Online]. Available: <http://science.sciencemag.org/content/359/6380/1146>
- [24] S. Wojcik, S. Messing, A. Smith, L. Rainie, and P. Hitlin, "Bots in the twittersphere," <http://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/>, 2018.
- [25] S. C. Woolley, "Automating power: Social bot interference in global politics," *First Monday*, vol. 21, no. 4, 2016.
- [26] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *ICDE*. IEEE, 2015.
- [27] L. Wu, F. Morstatter, X. Hu, and H. Liu, "Mining misinformation in social media," *Big Data in Complex and Social Networks*, 2016.
- [28] A. Zhang, B. Culbertson, and P. Paritosh, "Characterizing online discussion using coarse discourse sequences," in *ICWSM*, 2017.
- [29] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of WWW*, 2015.