Quantifying Deception Propagation on Social Networks

Information diffusion online is critical in many situations such as during crisis events, relaying health messages, or sharing news. As users rely on social media as primary sources of news and information, the propagation of deceptive information online is a significant concern. Recent studies have utilized linguistic models to prioritize content to fact-check (Gencheva et al., 2017), to distinguish trusted and deceptive content (Volkova et al., 2017), or to classify different types of suspicious content (Rashkin et al., 2017). We investigate the differences in the propagation of close to 11 million messages between January 2016 and January 2017 that retweet or mention 282 sources that have been identified as spreading verified (V) or suspicious news. We further classify suspicious news as clickbait (CB), conspiracy theories (CS), propaganda (P), or disinformation (D).

First, we looked at how propagation differs by demographics of users spreading the information. Identifying 145,688 users with at least five English-posts that reference deceptive accounts within our dataset, we use the Humanizr classifier (McCorriston et al., 2015), we classify each user as an organization or person. We then build Convolutional Neural Network (CNN) models initialized with GLoVe embeddings to infer the demographics of the 66,171 person-users who had a sufficient amount of publically available tweets. Looking at binary likelihoods of spreading information from a given type of source at least once over the collection period, we see statistically significant differences between classes of several demographics using Mann Whitney U tests. Women, on average, have a higher probability than men to have spread news from verified-, clickbait-, conspiracy-, and propaganda-sources ($p < .01$) but lower than men for disinformation-sources ($p = 0.01$). Single users are more likely ($p < .01$) of spreading information for all types of sources except for clickbait and disinformation, where single users are no more or less likely to propagate at least one post referencing those sources than those that are not single. Users who are 25 or older are more likely than those under 25 to share clickbait and propaganda but less likely to spread verified, conspiracy, or disinformation.

Overall, we see diffusion inequality across all types with a subset of users are responsible for a large proportion of each sources' retweets. Figure 1 illustrates how these inequalities differ across each types, compared to perfect equality or inequality and Table 1 shows the same using measures commonly used to describe income inequality. The Gini coefficient measures inequality in frequency distributions as a score ranging from 0 (perfect equality) to 1 (perfect inequality). The Palma ratio is the ratio of the shares of the top 10% in the frequency distribution to the bottom 40%; in perfect equality, the top 10% is responsible for one fourth of what bottom 40% is, resulting in a ratio of 0.25. Disinformation has the greatest inequality in retweet-diffusion where the 10% most prolific retweeting users who are referencing disinformation-spreading sources author 14.77 times as many tweets as the bottom 40%. Interestingly, clickbait and propaganda have less diffusion inequality than verified but this may be due to the sample sizes of these types within our dataset.

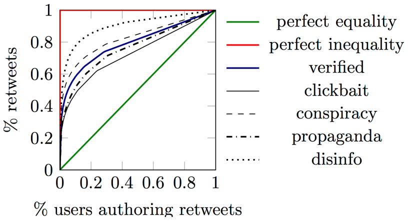

Figure 1. CDF of retweet diffusion as a function of the proportion of users authoring the retweets, for each type.

Table 1. Inequality measures for each message-type.

	V	*CB*	*CS*	*P*	*D*
Gini	0.59	0.46	0.65	0.53	0.80
Palma	4.01	2.39	5.35	3.14	14.77

While we use Twitter data, our network analyses are generalizable and can be used with other networks that utilize similar mechanisms for propagating information. Further our work is not message specific. Previous studies have utilized hashtags (Matsubara 2012) or common content (Kwon 2013) to identify individual events, topics, or rumors with which to compare propagation. We utilize sources rather than individual events or topics which eliminates any domain-specific event or topic tagging, relying instead on tagging through pre-existing links (retweets or mentions). This generalizability also allows us to capture the propagation behavior of the sources when propagating verified or non-deceptive messages as well as deceptive messages.

1. Gencheva, Pepa, et. al. "A Context-Aware approach for Detecting Worth-Checking Claims in Political Debates." RANLP, 2017.
2. Kwon, Sejeong, et al. "Prominent features of rumor propagation in online social media." *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013.
3. Matsubara, Yasuko, et al. "Rise and fall patterns of information diffusion: model and implications." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.
4. McCorriston, James, et al. "Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter." *ICWSM*. 2015.
5. Rashkin, Hannah, et al. "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking." *EMNLP*. 2017.
6. Volkova, Svitlana, et al. "Inferring Latent User Properties from Texts Published in Social Media." *AAAI*. 2015.
7. Volkova, Svitlana, et al. "Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter." *ACL* 2017.