Lost in diffusion: How (mis)information is distorted in online networks

Keywords: Misinformation, information diffusion, social network, information distortion, information ecosystem

Extended Abstract

The spread of misinformation has fueled the public's political polarization, vaccine hesitancy, and scientific uncertainties (Scheufele et al., 2021). Misinformation has received much academic attention, and the research foci have been on content identification, verification, and corrections (Shen et al., 2021). However, the dynamic evolution of misinformation and information flow has been overlooked.

Few studies examined how message content has been distorted, adapted, and deleted as messages spread over time. For example, Moussaid et al. (2015) found that messages tend to become shorter, inaccurate, and dissimilar when diffused in a chain, but judgment biases tend to become more extreme. A recent study found that misinformation messages changed aggressively at the initial stage of diffusion, and the sentiment and emotionality affected how they changed over time (King et al., 2021). Although those studies provide important insights, they overlooked characteristics that misinformation ignores the complexity of reality and relies on simplified causation such as conspiracy theories to encourage people to interpret, make causal judgments, or even imagine (Introne et al., 2020), which may generate different temporal distortion patterns of misinformation and information flow. Not to mention the role of human memory and attitudes in information distortion (Schacter, 2012). The public with varying levels of media literacy and different beliefs and values plays an essential role in disseminating and (re)creating information in the (mis)information ecosystem.

Relying on a comprehensive and longitudinal Twitter dataset, we examine the misinformation cascades from 2006 to 2017, to address the following questions: 1) How is true and false information distorted in the diffusion process? 2) What message factors (e.g., information veracity, sentiment) influence the temporal patterns of distortion?

Methods

Data. We are obtaining access to the anonymized dataset from MIT Media Lab (Vosoughi et al., 2018), which contains approximately 126,000 information cascades including the original tweets and retweets from 2006 to 2017. Both the original tweets and retweets have already been labeled for veracity by six independent fact-checking organizations into true, false, and mixed. In addition, in the dataset, the topic of tweets has been identified, the retweet networks have been captured, and individual characteristics of diffusers have also been recorded.

Operationalizing distortion with a combination of computational and qualitative methods. To identify distortion we will measure the semantic similarities or differences between tweets using clustering techniques such as cosine similarity. Cosine similarity models a text document as a vector of terms, and the similarity calculates the cosine value between two documents' term vectors. In addition, we propose a human annotating approach to operationalize distortion. A tentative human coding manual is developed to define an occurrence of distortion when the retweet either 1) (partially) removes information from the original tweet; 2) keeps the original information with different interpretation; or 3) adds more information that is not present in the original tweets.

Analysis. To analyze the data, we consider the information cascades as fully observed retweet networks, so the Twitter users who authored the original tweet and the subsequent retweets are nodes in a network, and a directed edge represents the retweeting behavior from one user to another. While accounting for networks' topology, classical survival regression models can be adapted to estimate the rate of distortion and node/edge attributes associated with distorted transmission. Treating distortion as a recurring event, we can estimate the hazard ratio of distortion by using the Andersen-Gill model, as a function of various endogenous and exogenous factors, including the sentiment of the tweet, the veracity of the original tweet, the shortest path from the original tweet, and user characteristics (i.e., the number of followers, account age, etc.).

Furthermore, human-annotated data can provide a more nuanced understanding of the information cascades with distortion. First, we are able to associate types of distortion (specifically, removing information, different interpretation, and adding information) with the rate of distortion in the process of diffusion. Second, the Markov chain can be applied to discover the pattern of (mis)information distortion, including how (mis)information distortion patterns progress and evolves over time, and how distortion types are connected in a temporal manner. Lastly, we compare cascade characteristics (i.e., diffusion depth, maximum breadth, size, structural virality) across (mis)information distortion types, looking at how distortion may influence the information diffusion patterns.

Outlook

By examining the over-time distortion and cascade of true and false information messages, we can better understand the mechanisms of (mis)information evolution in the diffusion process. Extending this line of investigation will help model the interdependence of (mis)information evolution with people's knowledge, beliefs, and attitudes, leading to a nuanced understanding of how misinformation spreads and persists, and further shedding light on differentiating and combating misinformation from receivers' perspective.

References

Introne, J., Korsunska, A., Krsova, L., & Zhang, Z. (2020). Mapping the Narrative Ecosystem of Conspiracy Theories in Online Anti-Vaccination Discussions. *International Conference on Social Media and Society*, 184–192. https://doi.org/10.1145/3400806.3400828

King, K., Wang, B., & Escobari, D. (2021). Effects of Sentiments on the Morphing of Falsehoods and Correction Messages on Social Media. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2021.789

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, *112*(18), 5631–5636. https://doi.org/10.1073/pnas.1421883112

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist*, *67*(8), 603.

Scheufele, D. A., Hoffman, A. J., Neeley, L., & Reid, C. M. (2021). Misinformation about science in the public sphere. *Proceedings of the National Academy of Sciences*, *118*(15), e2104068118. https://doi.org/10.1073/pnas.2104068118

Shen, C., Kasra, M., & O'Brien, J. F. (2021). Research note: This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-72

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559