What makes a rumor popular? A case study of Cryptocurrency Rumors on Reddit and Twitter

Abstract for SBP 2018 Data Challenge

Dian Hu, David A. Broniatowski

Department of Engineering Management & Systems Engineering, The George Washington University, Washington, DC, 20052, USA
{hudian,broniatowski} @email.gwu.edu

1 Introduction

In recent years, rumors, unverified news and other types of misinformation have gradually become a common phenomenon online (Del Vicario et al., 2016). Misinformation has effects on political campaigns (Allcott & Gentzkow, 2017), public health communication (Broniatowski, Hilyard & Dredze, 2016) and modern markets (DiFonzo & Bordia, 2007).

In the year of 2017, the world has witnessed the dramatic increase in the market value of bitcoin and several other types of cryptocurrency. In the meantime, we have also seen the wide-spreading of rumors of cryptocurrency on social media. In several cases, some involved companies or institutions of the rumors had to make public announcements to deny some rumors because they have affected the decision of too many investors.

However, rumors with the same gist might be told and retold by many individuals before or even after being rebuked by reliable sources. Each variation of these rumors might have different shares, upvotes or comments. What makes one rumor more popular than another one even when they have very similar gist? Is the popularity of a rumor entirely random or only determined by the number of followers of the rumor-spreading agents? Does the internal structure or syntax of the rumor message play a role when determining the popularity?

Fuzzy-Trace Theory (FTT), a leading theory of decision making, predicted that a message with causal explanation of otherwise mysterious adverse events would be more popular on the internet (Reyna, 2008). The landscape model, another related theory, argued that people will more easily comprehend and recall a message with a clear causal structure statement (Van den Broek, 1990). Based on landscape model, we infer that such a message will be also more popular on the internet. These prior works have hinted that the popularity of a message might also be related to the internal causal structures of that message.

2 Dataset and Question

In this study, we will analyze social media data revolving around 4 major cryptocurrencies from May 1st, 2017 to May 1st, 2018. As per the requirement of the data challenge, we will collect data from two different sources, Reddit and Twitter. We plan to measure how each of these rumor-related hyperlinks is being shared and discussed on either platform. The challenge question we are answering is question 3: "How does disinformation spread within and across media." Moreover, we plan to answer a specific question within the scope of the challenge question: "Will the existence of a clear causal structure within a rumor message facilitate the popularity of the rumor."

3 Method, Designs and Variables:

We collected 100000 Reddit messages randomly sampled from May 1st, 2017 to May 1st, 2018 with at least one of these 5 keywords "bitcoin," "xrp," "ethereum," "litecoin," "cryptocurrency." We will use Latent Dirichlet Allocation, a Bayesian topic model, to automatically segment the messages into 50 topics. We will infer which topic is related to a major rumor based on the first author's annotation.

Next, we will collect several features including Reddit upvote, number of followers, and whether the message contains a hyperlink to an external blog or news site. We plan to measure the readability of each tweet using the textstat python package. Readability metrics will be then decomposed into three dimensions using a principal component analysis, corresponding to comprehensibility (e.g., reading grade level), verbatim features (e.g., number of letters, number of words, etc.), and number of sentences. A logistic regression model will be used to assess factors associated with the likelihood that a Reddit message is being upvoted at least once. Also, we will use multiple linear regression to examine the factors associated with a total number of upvotes of each message.

A similar analysis will be performed on Twitter data as well. In the twitter data, instead of using Reddit upvote as the dependent variables, we will use the number of retweets as the dependent variables.

4 References:

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. National Bureau of Economic Research.
2. Broniatowski, D. A., Hilyard, K. M., & Dredze, M. (2016). Effective vaccine communication during the disneyland measles outbreak. Vaccine, 34(28), 3225–3228.
3. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554–559.
4. DiFonzo, N., & Bordia, P. (2007). Rumor psychology: Social and organizational approaches. American Psychological Association.
5. Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical Decision Making*, *28*(6), 850–865.
6. van den Broek, P. (1990). Causal inferences and the comprehension of narrative texts. *Psychology of learning and motivation*, *25*, 175–196.