Exploring the content of misinformation from multiple perspectives

Dian Hu, Advisor: David A. Broniatowski
Department of Engineering Management & Systems Engineering
The George Washington University

SBP-BRiMS 2017 Data Challenge

Abstract

Purpose:
In this data challenge project, we are interested in how the internal content of misinformation would affect online sharing behavior. In particular, we want to understand which internal factors of an article, if any, are predictive of how often the article is shared on Facebook. We explored three sets of independent variables.

Methods:
1. Our approach is to use regression analysis to explore which independent variables, are predictive of the number of Facebook shares of articles.
2. **Dependent Variables**: We gather Facebook share statistics using the Graph API provided by Facebook. Perform causal network analysis on a small set of randomly selected pilot data by treating rumors as narratives,
3. **Controlled Variables:** Prior studies have found several factors are significantly associated with information cascades [1][2]. These factors include, but are not limited to: (1) the popularity of the article's topic at the time of its sharing; (2) the article's length; (3) the popularity of the news platform. **To control these known significant factors respectively**, (1) we select news articles about a fixed topic: vaccination; (2) we only consider news published on and after Jan 1st, 2016, and (3) we analyze data from each news site separately.
4. **Independent Variables**:
   a) Five Measurements using Moral Foundations Dictionary (MFD).
   b) Readability Score, Article Length, and LSA Overlap Score.
   c) Network properties extracted from the network collected using the crowdsourcing tool we presented in the conference [3].

Data Explored:
- As per the requirement of the data challenge, we will analyze news data from the GDELT database gathered using the keyword "**vaccine**". We will also explore news data from a major anti-vaccine news website from which we have previously gathered data. From each news site, we collect 20 vaccine related articles published on and after Jan 1st, 2016.

Preliminary Results:
1. When looking into five dimensions of Moral Foundations Dictionary, we find that within each news media, each dimension exhibits different statistical power predicting Facebook shares.
2. We find no evidence showing that Readability Score, Article Length, and LSA Overlap Score are predictive of Facebook shares in this data set.
3. We find that although our crowdsourcing tool are effective to gather perceived causal network quickly, the interrater reliability of the response are still low. Therefore, we cannot confirm the network gathered are representative of how the structure would be perceived by the public at this moment.

Implication:
- Inspired on the interesting correlations emerged from the MFD dimensions, we plan to test whether the same pattern can be observed across news of various topics within each site; we also plan to test whether that pattern had changed from time to time within each site.
- We have extended our tool from studying simple narratives into studying news articles. However, much work is still needed to increase interrater reliability.

Representing Narratives as Causal Networks

• Prior studies of narrative have concluded that causal networks are a replicable way to represent how humans understand text [4]. When describing a narrative as a causal network, each event in the narrative is represented by a node, whereas each causal relationship between events is represented as a link [4][5].

• When readers draw the causal connections between events, they are showing us how they comprehend the text. Therefore, these responses are valuable for our research.

Event Index	Text
1	Daniel arrived in his aunt's house.
2	He knocked at the door and
3	rang the doorbell
4	Daniel's cousin was waiting at home.
5	He opened the door for Daniel then
6	two dogs ran out to greet Daniel.

Table 1. Labeled events in a simple narrative

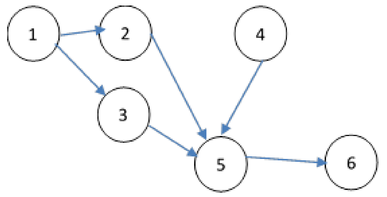

Fig. 1. Causal network associated with narrative presented in Table 1.

Preliminary Results: Networks From Crowdsourcing

- In this pilot study, we asked Mturk Workers to construct the causal networks of 6 selected news. For each news, 10 workers are recruited. Therefore, for each link between Node A to Node B, when it comes to how many people agree there is a link, there can be 11 possible outcomes (X=0, 1, 2 .. 10). We can present the aggregated graph by mandating that each pair must have more than N people agree. Based by literature in psycholinguistics [4][5], narrative's density is related to its internal coherence and therefore might affect people's acceptance of the content. Therefore, we aim to test the hypothesis whether the average *betweenness centrality* and the *total number of edge* of the graph is correlated to the Facebook shares.

- When presenting the aggregated network from the group, different N can lead to different result. If the graph change dramatically when N change, that means our interrater reliability is low. The following 3 graphs show how the aggregated graph of a news can change when N is 1,2 and 5. We need improve our crowdsourcing instructions to promote reliability and productivity of the crowdsourcing workers in the future in order to test the theory-based hypothesis.

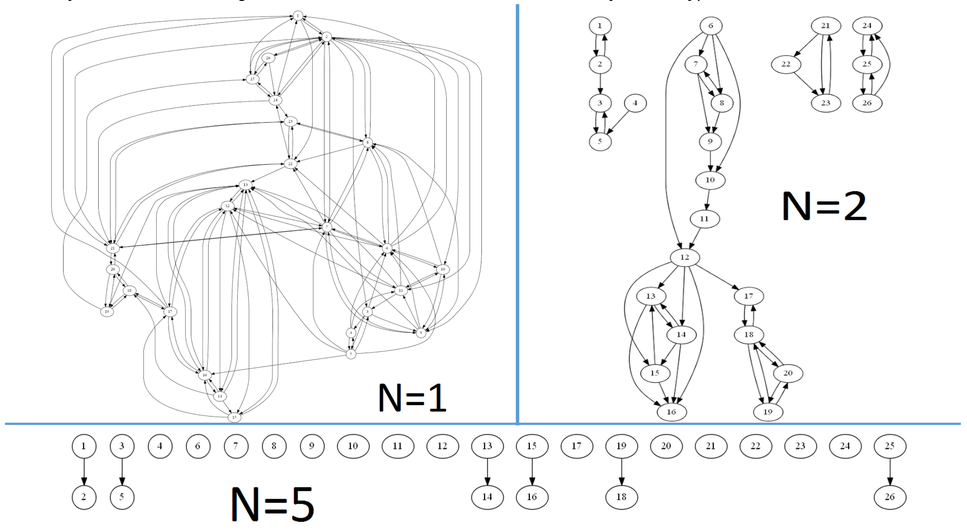

Fig. 2. Aggregated Causal Networks from crowdsourcing workers with different cutoff boundary

Preliminary Results: MFD and Facebook shares

•The Moral Foundations Dictionary is a dictionary of linguistic tokens derived from the study of Moral Foundations Theory [6][7]. The dictionary highlights five moral dimensions: **Authority, Fairness, Harm, Loyalty,** and **Purity.**

•For example, the words: "obey", "disobey", "comply", "oppose" and others are considered within the dimension of **Authority**. And "sacred", "defile", "innocent", "sin" and others are considered within the dimension of **Purity**. To reduce confounding effect, the word "disease" is removed from the original list during this project because we choose the topic of "vaccine" during data collection.

•We use normalized frequency of stem-match as an approximation of the magnitude of each moral dimension in each message.

• We found that on News Site 1, Purity dimension shows a positive trend towards the Facebook shares of that article (P=0.06, n=20), but barely missed P value of significance. **In another word, the viewers of News Site 1 are more likely to share vaccine news if the news use more Purity Tokens in the writing. However, on News Site 2, Purity dimension is not only far from being statistically significant, it's also show negative correlation. And authority dimension is showing strong negative correlation** (P=0.029, n=20).

• Nevertheless, this two findings are based on relatively small sample size (n=20). We must conduct a series of rigorous quasi-experiments to test whether the same pattern shown is true in different topics, different time period and different sample size.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.355950859	0.557639189	6.01814	3.155E-05
Harm	-37.09765502	21.95755121	-1.68952	0.1132566
Purity	49.9589669	24.5026102	2.038924	0.0608029
Ingroup	-57.81018002	32.61589543	-1.77245	0.09807
Fairness	18.41172779	21.02858134	0.875557	0.3960445
Authority	-18.11252838	11.01294193	-1.64466	0.1222988

Table 2: Linear Regression: MFD against Facebook Share, Site 1

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.646596795	0.371184787	9.824209	1.163E-07
Harm	42.39885173	28.35837844	1.495108	0.1570833
Purity	-33.35921399	31.2458577	-1.06764	0.3037494
Ingroup	21.25483648	17.52070619	1.213127	0.2451536
Fairness	-76.74820846	64.93795878	-1.18187	0.2569437
Authority	-54.65509036	22.55361202	-2.42334	0.0295179

Table 2: Linear Regression: MFD against Facebook Share, Site 2

References

[1] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, pp. 80–88, 2010.

[2] J. M. Hofman, A. Sharma, and D. J. Watts, "Prediction and explanation in social systems," Science, vol. 355, no. 6324, pp. 486–488, 2017.

[3] D. Hu and D. A. Broniatowski, "Measuring Perceived Causal Relationships Between Narrative Events with a Crowdsourcing Application on Mturk," in International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, 2017, pp. 349–355.

[4] T. Trabasso and others, "Causal cohesion and story coherence.," 1982.

[5] T. Trabasso and P. Van Den Broek, "Causal thinking and the representation of narrative events," Journal of memory and language, vol. 24, no. 5, pp. 612–630, 1985.

[6] J. Graham et al., "Moral foundations theory: The pragmatic validity of moral pluralism," 2012.

[7] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations.," Journal of personality and social psychology, vol. 96, no. 5, p. 1029, 2009.