News Diffusion Through Online Social Network

Jingying (Jane) Bi | Email: jingyingb@uchicago.edu | Master of Arts Program in the Social Sciences (MAPSS), University of Chicago

Online Social Network

ABSTRACT

False news diffuses unprecedentedly faster through the online social networks nowadays. It causes unnecessary chaos or panic and might even harm the democracy. I collected the Tweets containing both *true* and *false* news via Twitter APIs, analyzed and compared their diffusion patterns, as well as studied the impact of the engaged user's characteristics on the news diffusion. I find that the *false* news diffuses significantly faster, deeper and broader than the *true* news. The news also spreads wider if it circulates among users with less followers but more friends, posting less Tweets but tending to like more other users' posts. Furthermore, Twitter community could recognize the *false* news and try to correct/filter it over time. Three distinct false-correction feature are identified. Examining them might help discover and curb the *false* news.

INTRODUCTION

News Content	News Veracity	# Tweets collected	# Users Engaged
HIV-infected blood was injected into Pepsi or Frooti plant.	False	6,113	5,440
Type "BFF". If it turns green, your Facebook is secure.	False	964	911
Coors Light beer was contaminated with cocaine.	False	578	568
Coca-Cola recalled Dasani because a clear parasite was found in bottles across the United States.	False	1,558	1,515
Emma González, a Parkland mass shooting survivor, admitted to bullying a former student who killed 17 people.	False	9,007	8,441
Soros paid $300 to 'March for Our Lives' Protesters.	False	807	762
Facebook categorizes users by political preferences	True	1,383	1,358
50 million Facebook profiles data leaked out.	True	24,771	20,919

METHODS

STEP I: The data is crawled via the Application Programming Interface (API), including *Twarc, OldTweet*, and *Tweepy.* The data collection Python codes are run on the RCC Midway.

STEP II: Label the Tweets as "true", "false", "correction", or "irrelevant". Recover the news diffusion process, i.e. the Retweets topology, based on the two assumptions.

1) A user retweets a news at the first time she reads a Tweet containing that news.
2) A user only reads the Tweets posted by the users that he follows, i.e. his friends.

STEP III: Let T_i represent Tweet i. RT_{ij} is the jth Retweet of T_i. T_user_i posted T_i. RT_user_{ij} posted RT_{ij}. $Friends_{ij}$ is the set of RT_user_{ij}'s friends. $Earlier_{ij}$ is the subset of $Friends_{ij}$ who also retweeted T_i but at an earlier time than RT_user_{ij}. Clearly, T_i is the parent of $RT_{ij}, \forall j$. Among all the RT_{ij}, the *parent-child* relationship is configured by the following algorithm. The process is shown in **Fig. 1**.

1) If $Friends_{ij} \cap Earlier_{ij} = \emptyset$, then RT_{ij} is directly retweeted from T_i.
2) Otherwise, among the users in $Friends_{ij} \cap Earlier_{ij}$, for the user RT_user_{ij*} who retweeted T_i latest, his post RT_{ij*} is the parent of RT_{ij}.

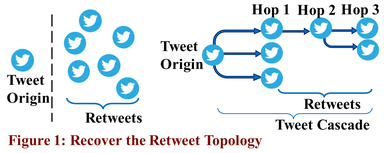


Figure 1: Recover the Retweet Topology

STEP IV:

Linear Regression Model

$$Pattern_i = \beta_0 + \beta_1 Veracity_i + \beta_2 X_i + \beta_3 U_i + \varepsilon_i$$

Logistic Regression Model

$$Veracity_i = \beta_0 + \beta_1 Depth_i + \beta_2 Speed_i + \beta_3 Breadth_i + \beta_4 X_i + \beta_5 U_i + \varepsilon_i$$

Where *Pattern* includes three characteristics, namely speed, depth, and breadth. Veracity is a binary variable of "True", "False", and "Correction" levels. X gives more information about the Tweets. U gives more information of the engaged users.

RESULTS

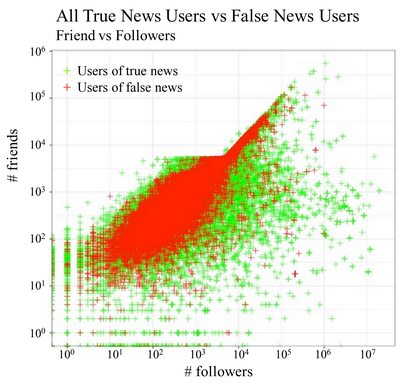


Figure 2: Number of Friends vs Followers

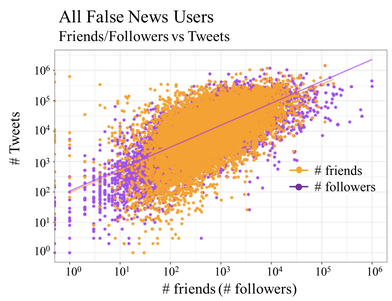


Figure 3: Number of Tweets vs Friends/Followers

Fig. 2 displays a flat cut at 5,000 friends level. This is due to Twitter's following limit. Each Twitter user could follow at most 5,000 accounts. Once the limit is reached, the user has to wait for more people to follow him, until he could continue to have more friends. For those accounts with more than 5,000 followers, this limit does not apply, and hence it is possible to get more friends than 5,000. **Fig. 3** displays a positive relationship between the number of Tweets and the number of friends or followers.

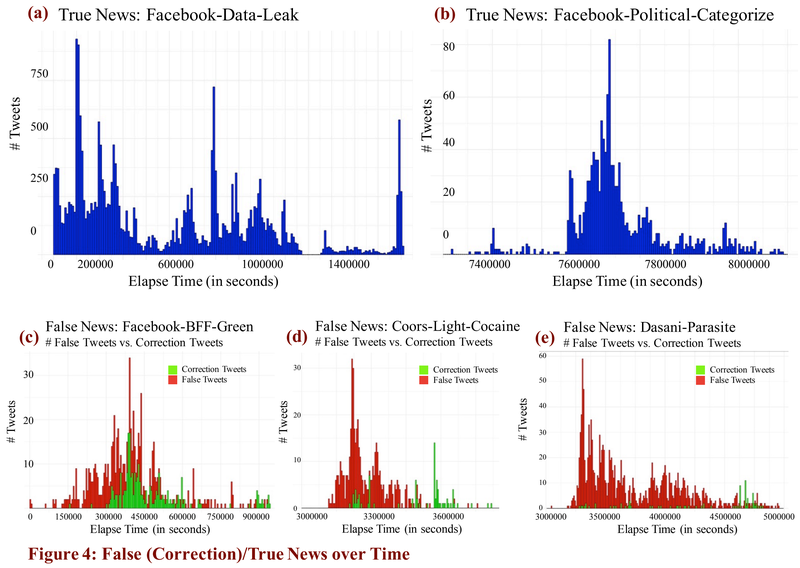


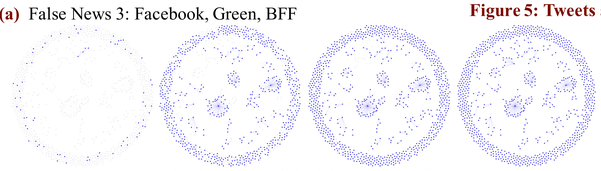
Figure 4: False (Correction)/True News over Time

Fig. 4 panels **(a)** and **(b)** are the true news distributions over time. Panels (c)~(e) display three false-correction features. Panel **(c)** gives the *concurrent* feature. The distribution of the correction Tweets is the same as that of the false Tweets, but starting later and to a smaller extent. Panel **(d)** shows the *gothic* feature. Specifically, the distribution of false Tweets looks thin and high as the Gothic buildings. There are only a few correction Tweets which emerged rather late. This may be due to the outbreak of the false news which also cooled down shortly. Thus, the diffusion peak was reached too fast to have the correction Tweets being composed and propagated. By the time of the correction Tweets started to circulate, the false news no longer caught much public attention. Consequently, the correction Tweets dwindled soon as well. Panel **(e)** shows the *camel* feature. The false Tweets diffused more than one round, where the distribution looks like the humps of the camel. the first peak was reached shortly, then followed by a downslope, and then a second peak. There are little corrections happening during this period. More correction Tweets emerged at the end of the second "hump" and tended to outweigh the false news seeking for a third but dwarf peak.

RESULTS

(a) False News 3: Facebook, Green, BFF

Figure 5: Tweets and Retweets Topology



(b) True News 10: Facebook Users, Political Category, Preference

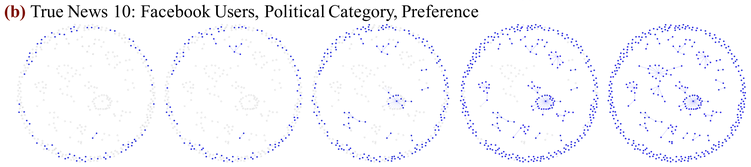


Table 1: Percentage of nodes by each period

News Id	By period 1	By period 2	By period 3	By period 4	By period 5
3 (False)	7.94%	73.34%	94.93%	100%	--
4 (False)	54.24%	86.22%	97.53%	100%	--
6 (False)	46.25%	74.3%	92.57%	100%	--
8 (False)	59.78%	74.38%	82.61%	100%	--
10 (False)	19.84%	28.31%	49.74%	85.19%	100%

Table 2: Linear/Logistic Regression Results

	Dependent variable: Width			Dependent variable: Veracity	
	Base	+ User Info	Full	Base	Full
	(1)	(3)	(4)	(1)	(3)
Label "Correction"	-370.225*** (18.991)	-367.28*** (18.939)	-671.60*** (12.128)		
Label "False"	-503.688*** (15.412)	-507.35*** (15.406)	-650.68*** (9.529)		
Time elapse			0.0001*** (0.00001)		
Depth			28.673*** (4.096)		
Speed_2			40,425.7*** (2,332.514)		
Breadth			1.388*** (0.023)		
Tweet origin			-0.304 (9.102)		
User favorite no.		0.001*** (0.0002)	0.0004*** (0.0001)		
User friend no.		0.001 (0.001)	0.002*** (0.001)		
User follower no.		-0.001*** (0.0001)	-0.0005*** (0.0001)		
User Tweet post no.		-0.0001 (0.0001)	-0.0001*** (0.00004)		
Constant	801.694*** (14.283)	799.16*** (14.716)	558.18*** (13.102)		
Tweet Origin				2.167*** (0.291)	2.218*** (0.297)
Depth				0.57*** (0.182)	0.631*** (0.186)
Speed_2				227.24*** (49.545)	246.051*** (50.344)
Breadth				0.014*** (0.001)	0.014*** (0.001)
# Favorite					-1.093e-06 (2.055e-06)
# Friend					-7.583e-06 (1.105e-05)
# Follower					-5.332e-06*** (1.831e-06)
# Tweet Post					-2.599e-06** (1.157e-06)
Constant				-1.93*** (0.315)	-1.874*** (0.320)
Obs	2,331	2,331	2,331	1,691	1,691
R2	0.3187045	0.3347669	0.7841592	0.3206618	0.8313201
AIC	32,134.590	32,086.980	29,473.210	-17,690.31	18,456.04
Note:		*p<0.1; **p<0.05; ***p<0.01		*p<0.1; **p<0.05; ***p<0.01	

CONCLUSIONS

- False news diffuses faster, deeper and broader than the true news.
- News spreads wider (i.e. engage more Tweet origins) if it circulates among users with less followers but more friends, posting less Tweets but liking more other users' posts.
- Twitter community could recognize the false news and try to correct it over time. Three false-correction features are identified, namely *concurrent*, *gothic*, and *camel* features.