

CS6471 – COMPUTATIONAL SOCIAL SCIENCE

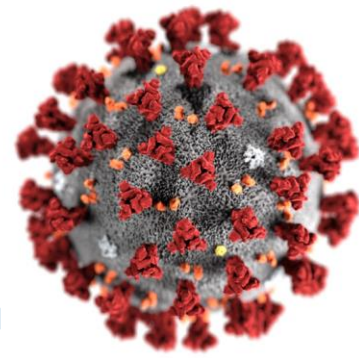
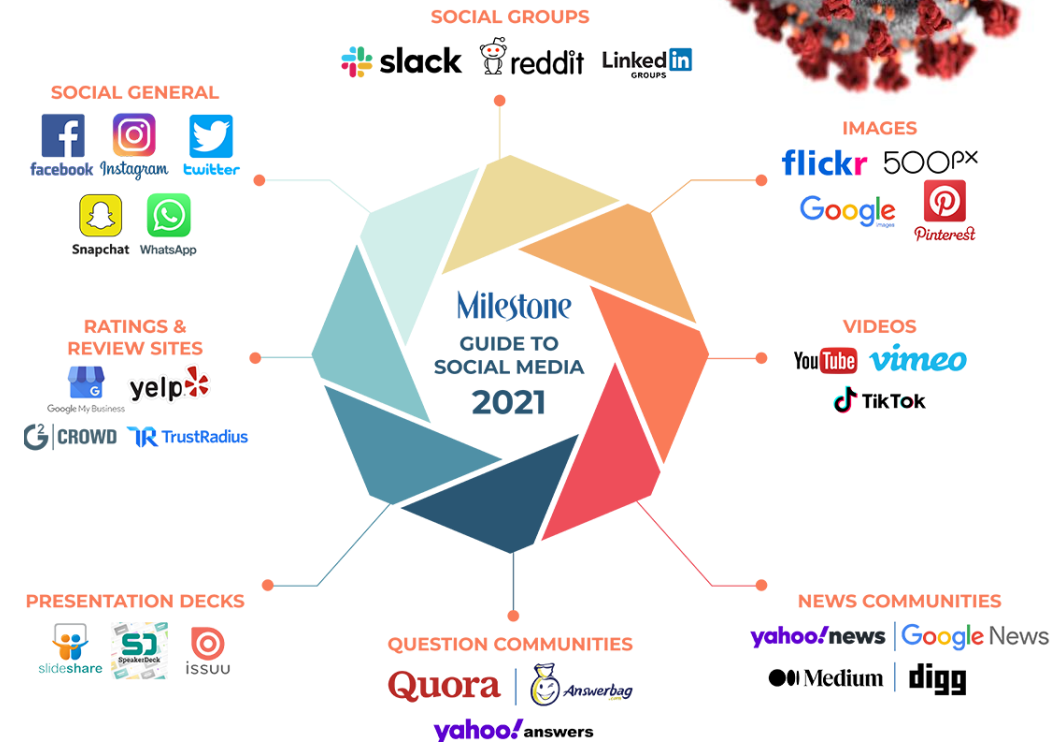
Information Diffusion: Misinformation

YOON SUN BYUN, WILSON PU,
ZHANZHAN ZHAO



Misinformation and Importance

- Misinformation is a like a “disease” and “resilient virus” (Brian X. Chen, NYT)
- Exists in inundated information marketplace in low-cost forms of blogs/videos/ tweets/memes/etc. and leads to proliferation of online information
- Inherent cognitive biases that evolutionarily served us well are amplified in harmful ways by modern technology (search engines, social media, bots – automated social media accounts)
- Negative effects include manipulating people, being a detriment to well-being, stoking anger, causing violence
- Mitigate negative effects through understanding where misinformation comes from, how to detect it, and how it spreads



CAREER FEATURE · 17 JUNE 2020 · CLARIFICATION 24 JUNE 2020

Coronavirus misinformation, and how scientists can help to fight it

Bogus remedies, myths and fake news about COVID-19 can cost lives. Here's how some scientists are fighting back.

Nic Fleming

Agenda

- “The Spread of True and False News Online” (35 min)
 - Overview
 - Importance
 - Methods & Analyses
 - Strengths & Weaknesses
 - Q&A/Discussion
- “The Limitations of Stylometry for Detecting Machine-Generated Fake News” (35 min)
 - Overview
 - Importance
 - Methods & Analyses
 - Strengths & Weaknesses
 - Q&A/Discussion

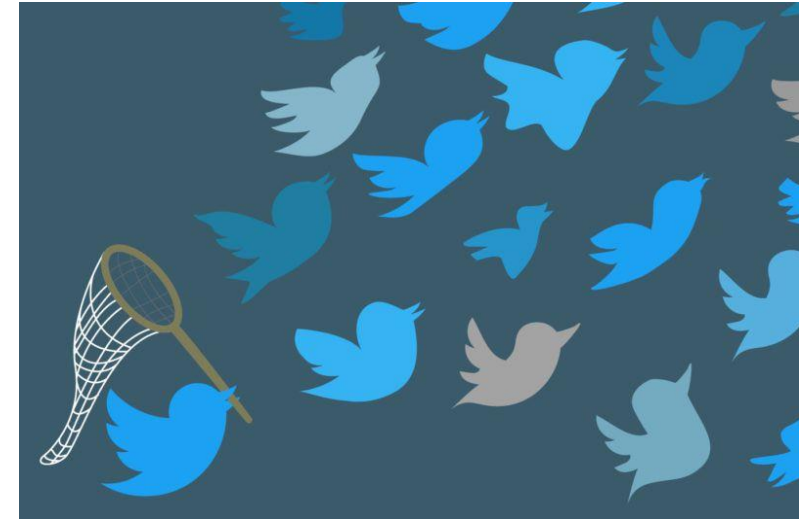
The Spread of True and False News Online

VOSOUGHI, S., ROY, D., & ARAL, S. (2018).
THE SPREAD OF TRUE AND FALSE NEWS
ONLINE. SCIENCE, 359(6380), 1146-1151.



Why is this topic important?

- The spread of falsehood is going viral
 - Further, Faster, Deeper, and Broader than the truth
 - Significant falsehood dispersion found in some areas like politics more than in other areas like terrorism, natural disasters, or finance
- Avoiding a fluid terminology "fake news"
 - Introducing more objectively verifiable terms
 - "True" and "False" news – attention on the veracity
- Need of analyzing the differential diffusion of "True" and "False" news stories
 - Examine why false news may spread differently than the truth



The Overview of the Paper



- Previous Work

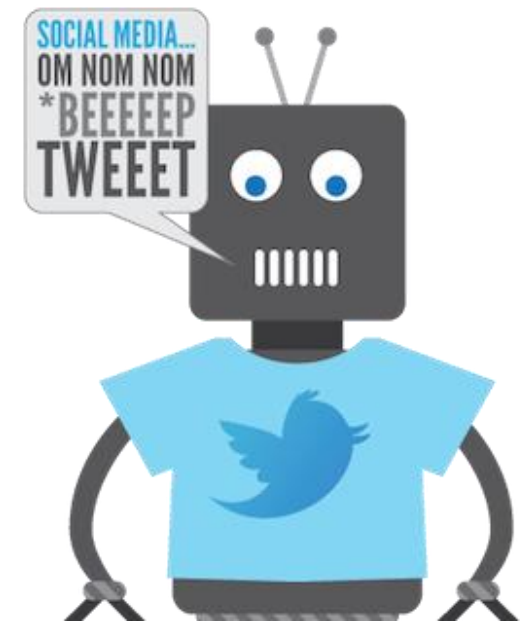
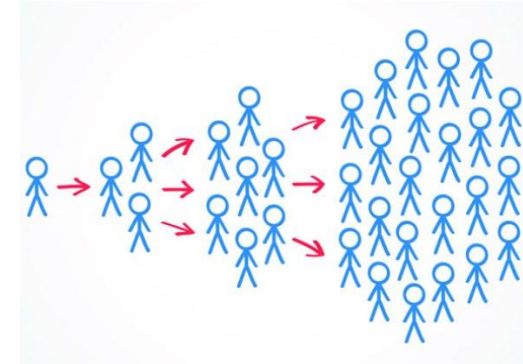
- Few large empirical studies of diffusion of misinformation or its social origins
- Fail to answer "how do truth and falsity diffuse differently?" & "what factors of human judgment explain these differences?"

- Approach

- Investigate differential diffusion of true, false, and mixed fact-checked news stories
 - Tweets from '06-'17 of ~126,000 rumor cascades spread by ~3 million people, ~4.5 million times
- Sampled all rumor cascades investigated by 6 fact-checking orgs and measured depth, size, maximum breadth, structural virality of cascade increase

- Insights:

- Bots accelerate true and false info at same rate
- Humans are the cause of the false information diffusing more than true



The background of the slide is a dense collection of small, colorful wooden human figures in various colors including orange, red, brown, and light wood. A white, torn-edge paper graphic is overlaid on the center, containing the title and list. The title is in a large, black, sans-serif font. The list items are in a smaller, black, sans-serif font, each preceded by a bullet point.

Methods & Analyses - Data

- Data: all of the verified true and false news stories distributed on Twitter from 2006 to 2017:
- -> ~126,000 stories
- -> stories tweeted by ~3 million people more than 4.5 million times
- -> classified as true or false by six independent fact-checking organizations
- -> those six organizations exhibited 95 to 98% agreement on the classifications.

Methods & Analyses - Terminologies

Rumors: news widely spread

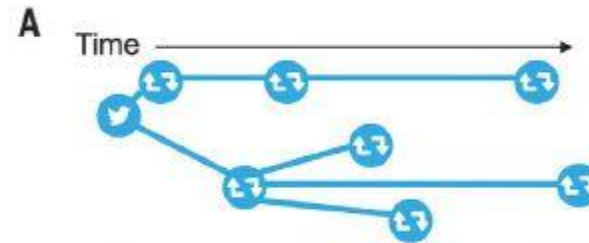
-> true rumors: true news

-> false rumors: false news

To quantify the diffusion dynamics of rumors:

-> cascade

-> cascade #

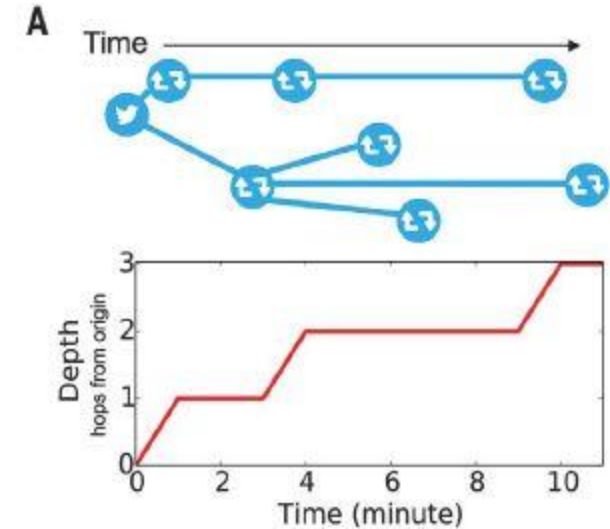


If a rumor "A" is tweeted by 10 people separately, but not retweeted, it would have 10 cascades, each of size one. Conversely, if a second rumor "B" is independently tweeted by two people and each of those two tweets is retweeted 100 times, the rumor would consist of two cascades, each of size 100.

Methods & Analyses - Terminologies

To quantify the diffusion dynamics of rumors:
-> cascades' depth

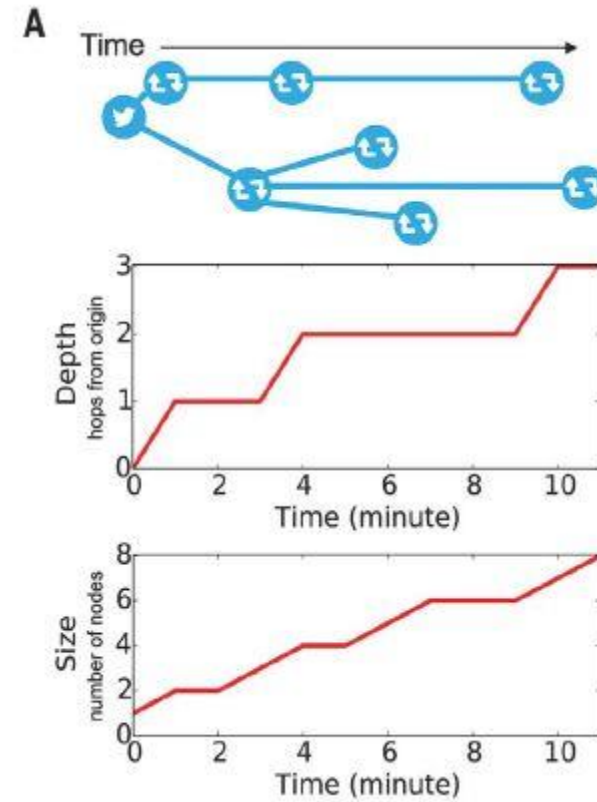
the number of retweet hops from the origin tweet over time, where a hop is a retweet by a new unique user.



Methods & Analyses - Terminologies

To quantify the diffusion dynamics of rumors:
-> cascades' size

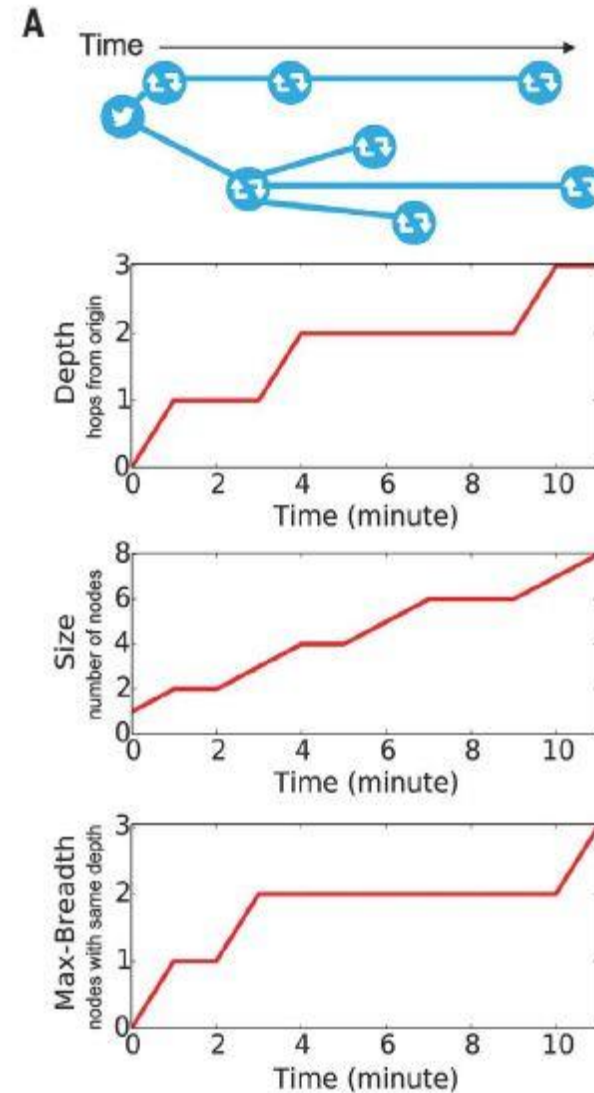
the number of users involved in the cascade over time.



Methods & Analyses - Terminologies

To quantify the diffusion dynamics of rumors:
-> cascades' maximum breadth

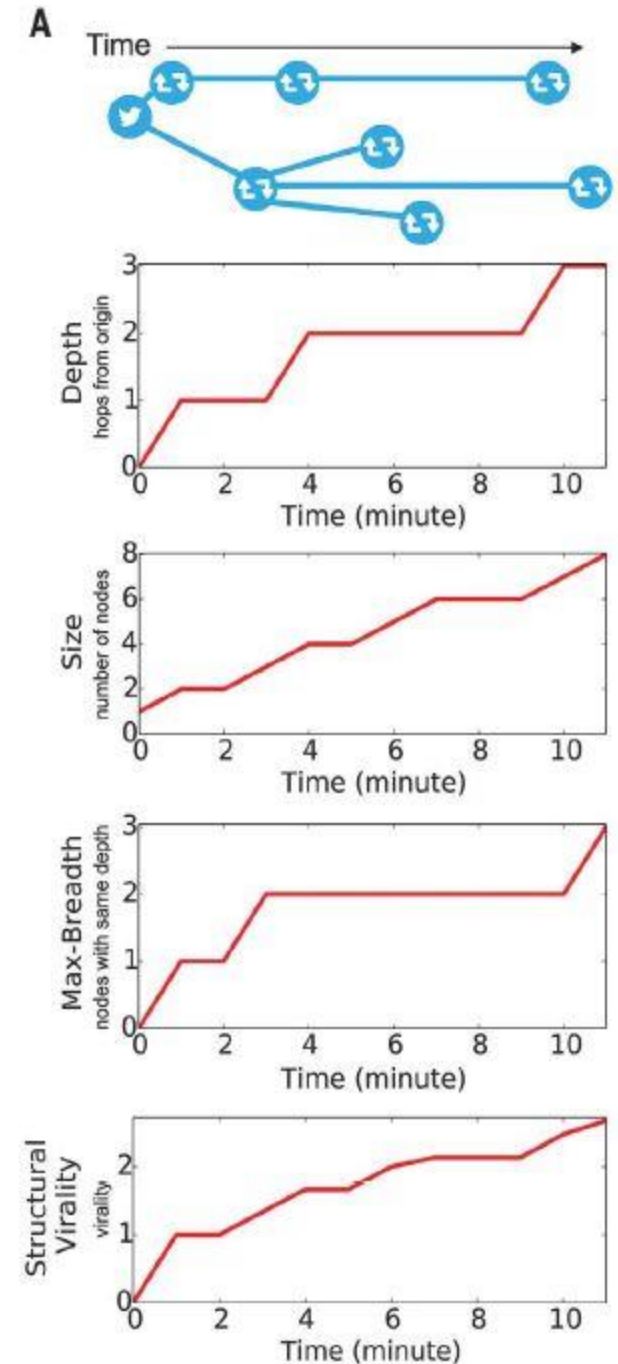
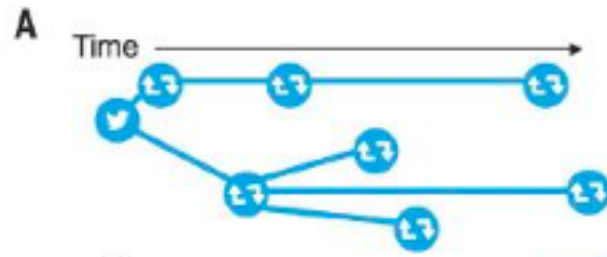
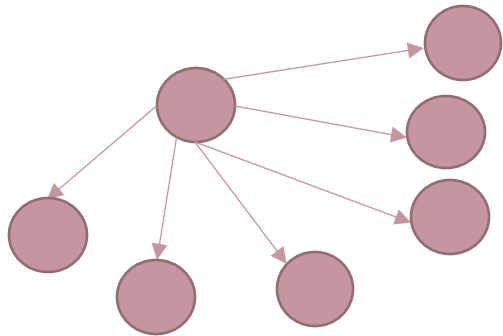
the maximum number of users involved in the cascade at any depth.



Methods & Analyses - Terminologies

To quantify the diffusion dynamics of rumors:
-> cascades' structural virality

a measure that interpolates between content spread through a single, large broadcast and that which spreads through multiple generations, with any one individual directly responsible for only a fraction of the total spread.

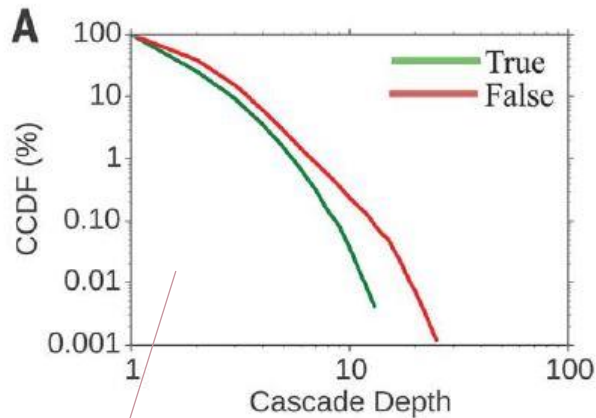




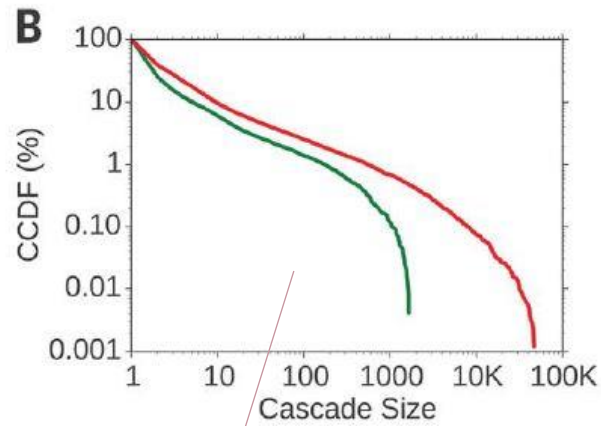
Methods & Analyses - *Diffusion Dynamics*

- Dynamics and comparison between the false news and true news
- Dynamics and comparison between political false news and all other category false news

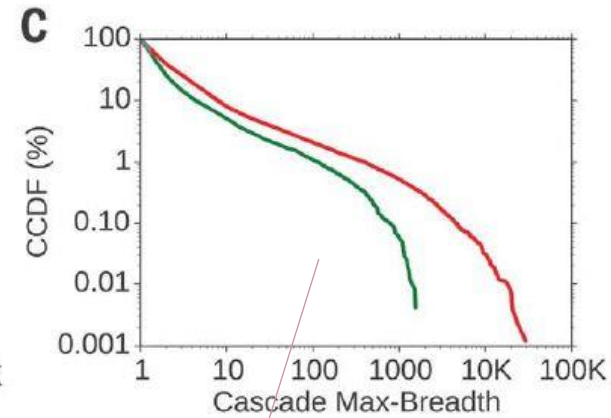
Methods & Analyses – Diffusion Dynamics



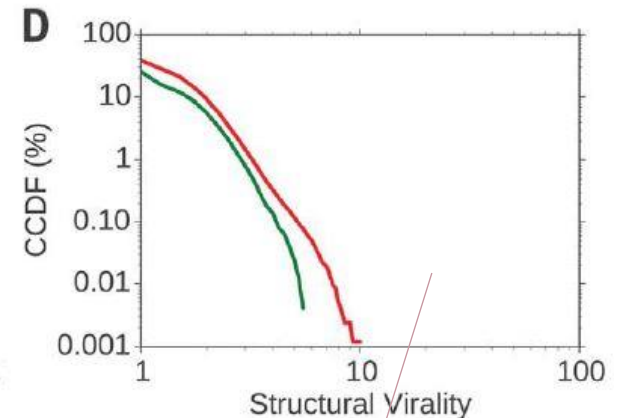
A significantly greater fraction of false cascades than true cascades exceeded a depth of 10, and the top 0.01% of false cascades diffused **eight** hops deeper into the Twittersphere than the truth, diffusing to depths greater than **19** hops from the origin tweet.



Whereas the truth rarely diffused to more than 1000 people, the **top 1%** of false-news cascades routinely diffused to between **1000 and 100,000 people**.

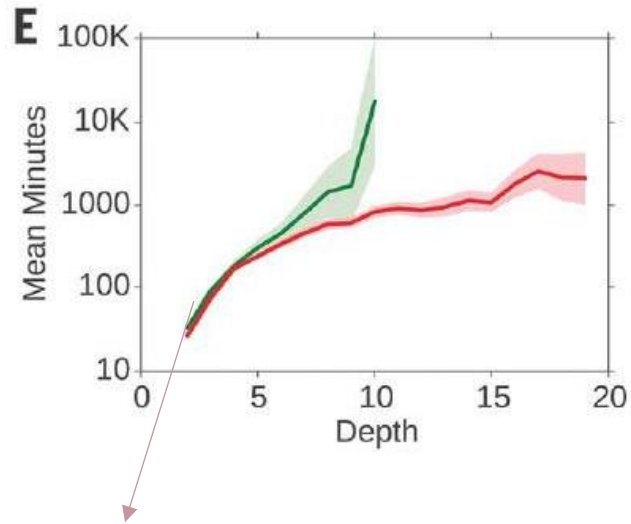


Falsehood reached more people at every depth of a cascade than the truth, **meaning that many more people retweeted falsehood than they did the truth**.

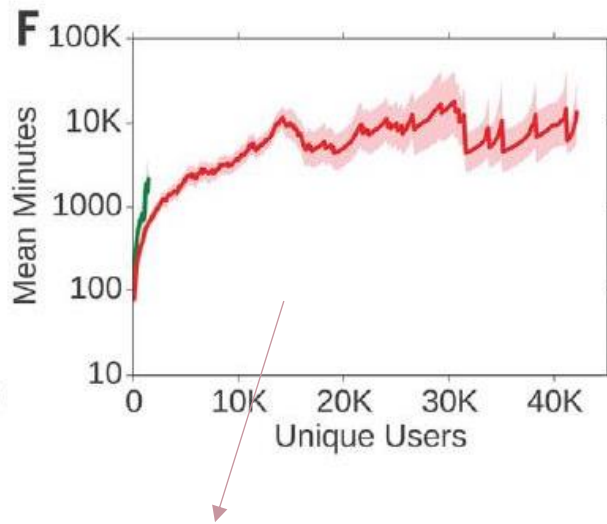


The spread of falsehood was aided by its virality, meaning that falsehood did not simply spread through broadcast dynamics but rather through **peer-to-peer diffusion** characterized by a viral branching process.

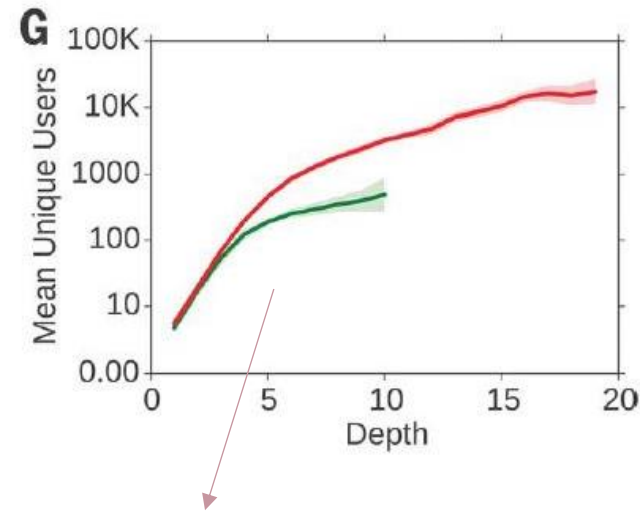
Methods & Analyses – Diffusion Dynamics



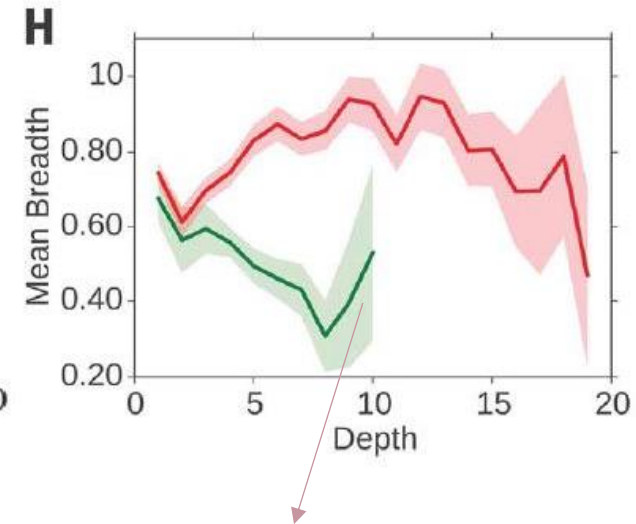
It took the truth about **20 times** as long as falsehood to reach a cascade **depth of 10**. As the **truth never diffused beyond a depth of 10**, we saw that falsehood reached a depth of **19** nearly **10 times faster** than the **truth** reached a depth of 10.



It took the truth about **six times** as long as falsehood to reach **1500 people**.

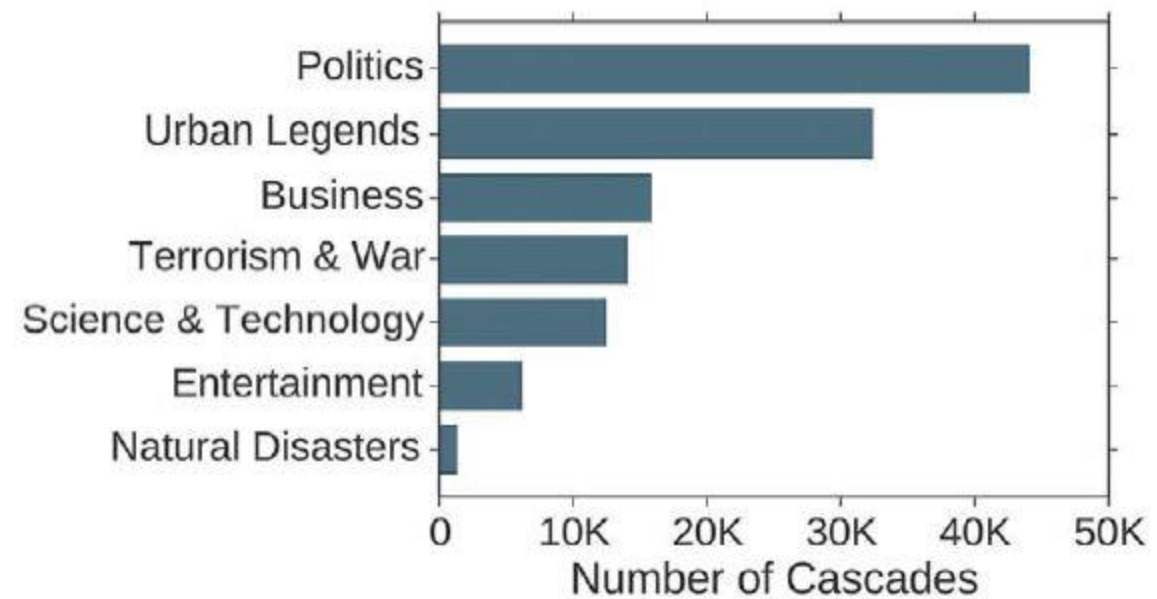


Falsehood was retweeted by **more unique users** than the truth at every cascade depth.



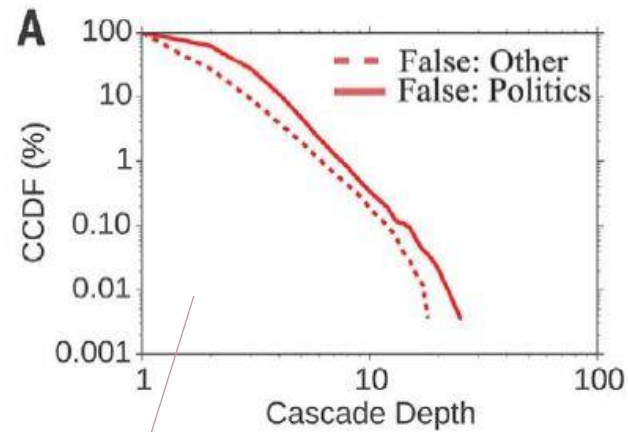
Falsehood also diffused significantly **more broadly**.

Methods & Analyses – Diffusion Dynamics

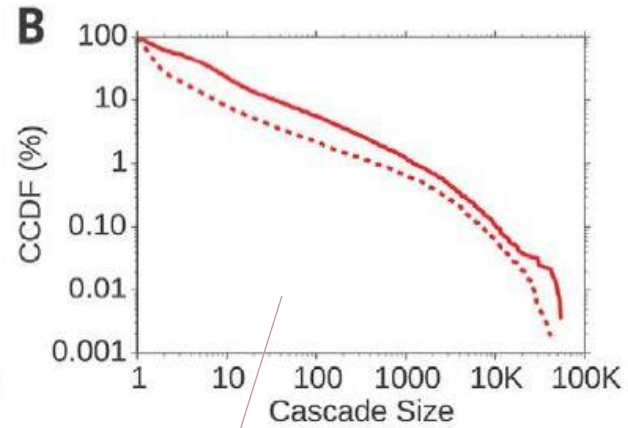


A histogram of the total number of rumor cascades in our data across the seven most frequent topical categories.

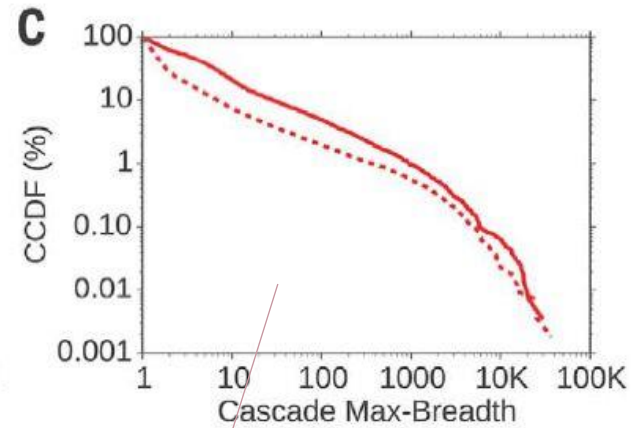
Methods & Analyses – Diffusion Dynamics



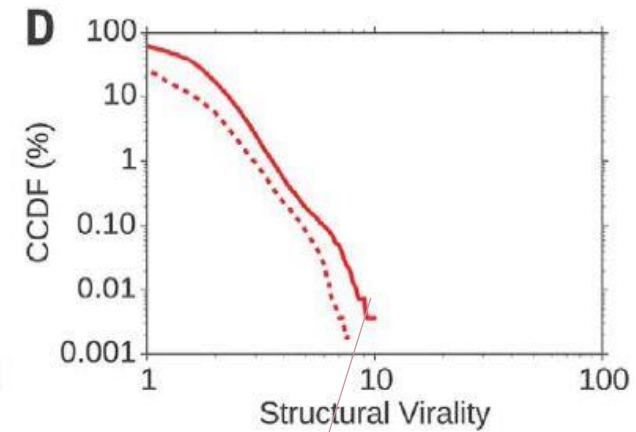
False political news traveled **deeper** than any other category of false information.



False political news **reached more people**.

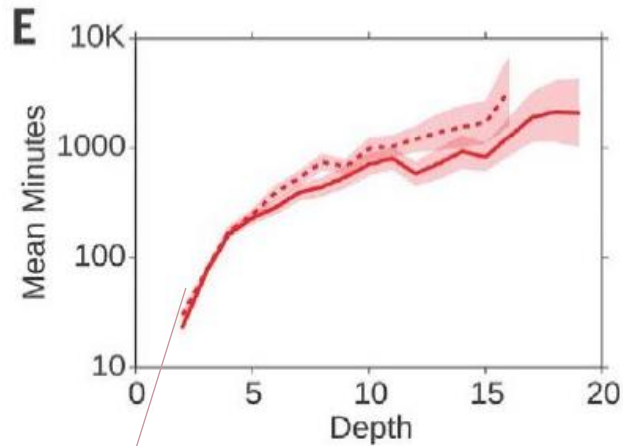


False political news traveled more **broadly**.

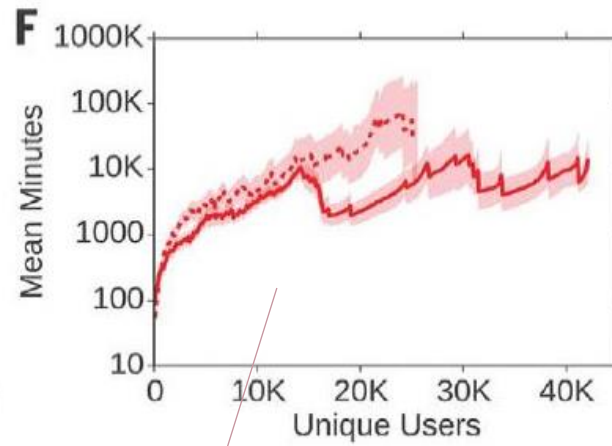


False political news was **more viral** than any other category of false information.

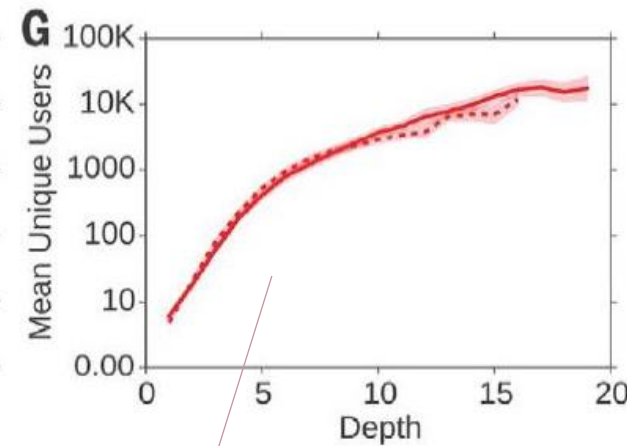
Methods & Analyses – Diffusion Dynamics



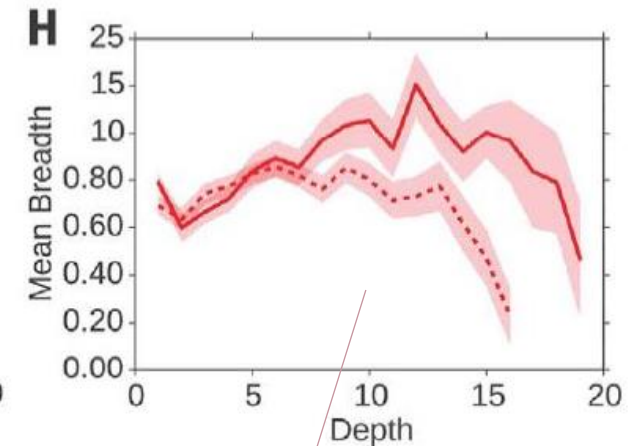
False political news diffused deeper more quickly.



False political news reached **more than 20,000** people nearly **three times** faster than all other types of false news reached **10,000 people**.



Although the other categories of false news reached about the same number of unique users at depths between 1 and 10, **false political news routinely reached the most unique users at depths greater than 10.**



Although all other categories of false news traveled slightly more broadly at shallower depths, false political news traveled more broadly at greater depths, indicating that **more-popular false political news** items exhibited broader and more-accelerated diffusion dynamics.

The background of the slide is a dense collection of small, colorful wooden human figures in various colors including orange, red, brown, and light wood. A white, torn-edge paper graphic is overlaid in the center, containing the title and list.

Methods & Analyses – Inferring False News Diffusion

- Why such false news diffusion?
- Users' characteristics?
- Network structure?
- Novelty of the false news?
- Users' perceptual emotions?

Methods & Analyses – *Inferring False News Diffusion*

- Users' characteristics?

Comparison of users involved in true and false rumor cascades:

Users who spread false news had significantly fewer followers, followed significantly fewer people, were significantly less active on Twitter, were verified significantly less often, and had been on Twitter for significantly less time.

Falsehood diffused farther and faster than the truth despite these differences, not because of them.

	median		mean		mean (log)		stdv (log)		ks-test
	false	true	false	true	false	true	false	true	
followers	410	466	2234	5240	2.62	2.68	0.69	0.88	D=0.104, p~0.0
followees	383	509	1002	1707	2.59	2.72	0.85	0.96	D=0.136, p~0.0
verified	0	0	0.002	0.006	nd	nd	nd	nd	D=0.005, p<0.001
engagement	9.52	9.54	19.70	24.65	0.91	0.90	0.65	0.76	D=0.054, p~0.0
account age	982	1214	1072	1269	2.90	2.97	0.39	0.42	D=0.125, p~0.0

Methods & Analyses – Inferring False News Diffusion

- Network structure?

Build a model of the likelihood of retweeting:

It is found that **falsehoods were 70% more likely to be retweeted than the truth even when controlling** for the account age, activity level, and number of followers and followees of the original tweeter, as well as whether the original tweeter was a verified user.

	coef	odds ratio	std err	z	P> z	[95.0% Conf. Int.]
account age	0.0002	1.000160	2.07e-05	7.759	0.000	0.000 0.000
engagement	0.0066	1.006648	0.000	18.019	0.000	0.006 0.007
falsehood	0.5350	1.707489	0.084	6.366	0.000	0.370 0.700
followees	-1.639e-05	0.999984	8.73e-06	-1.877	0.060	-3.35e-05 7.22e-07
followers	5.192e-05	1.000052	7.77e-06	6.682	0.000	3.67e-05 6.72e-05
intercept	-2.3941	0.091257	0.072	-33.437	0.000	-2.534 -2.254
verified	1.4261	4.162467	0.090	15.915	0.000	1.250 1.602

Methods & Analyses – *Inferring False News Diffusion*

- Novelty of the false news?

Measure how novel the information in the true and false rumors by comparing the topic distributions of the rumor tweets with the topic distributions of the tweets to which users were exposed in the 60 days before their retweet.

They found that false rumors were significantly more novel than the truth across all novelty metrics, displaying significantly higher information uniqueness.

	mean		variance		ks-test
	false	true	false	true	
IU	0.85	0.78	0.0052	0.0072	D=0.457, p~0.0
KL	4.49	4.15	0.1618	0.0948	D=0.433, p~0.0
BD	0.87	0.84	0.0008	0.0008	D=0.415, p~0.0

Data: Randomly selected ~5000 users who propagated true and false rumors and extracted a random sample of ~25,000 tweets that they were exposed to in the 60 days prior to their decision to retweet a rumor.

Then trained on 10 million English-language tweets, to calculate the information distance between the rumor tweets and all the prior tweets that users were exposed to before retweeting the rumor tweets.

Methods & Analyses – *Inferring False News Diffusion*

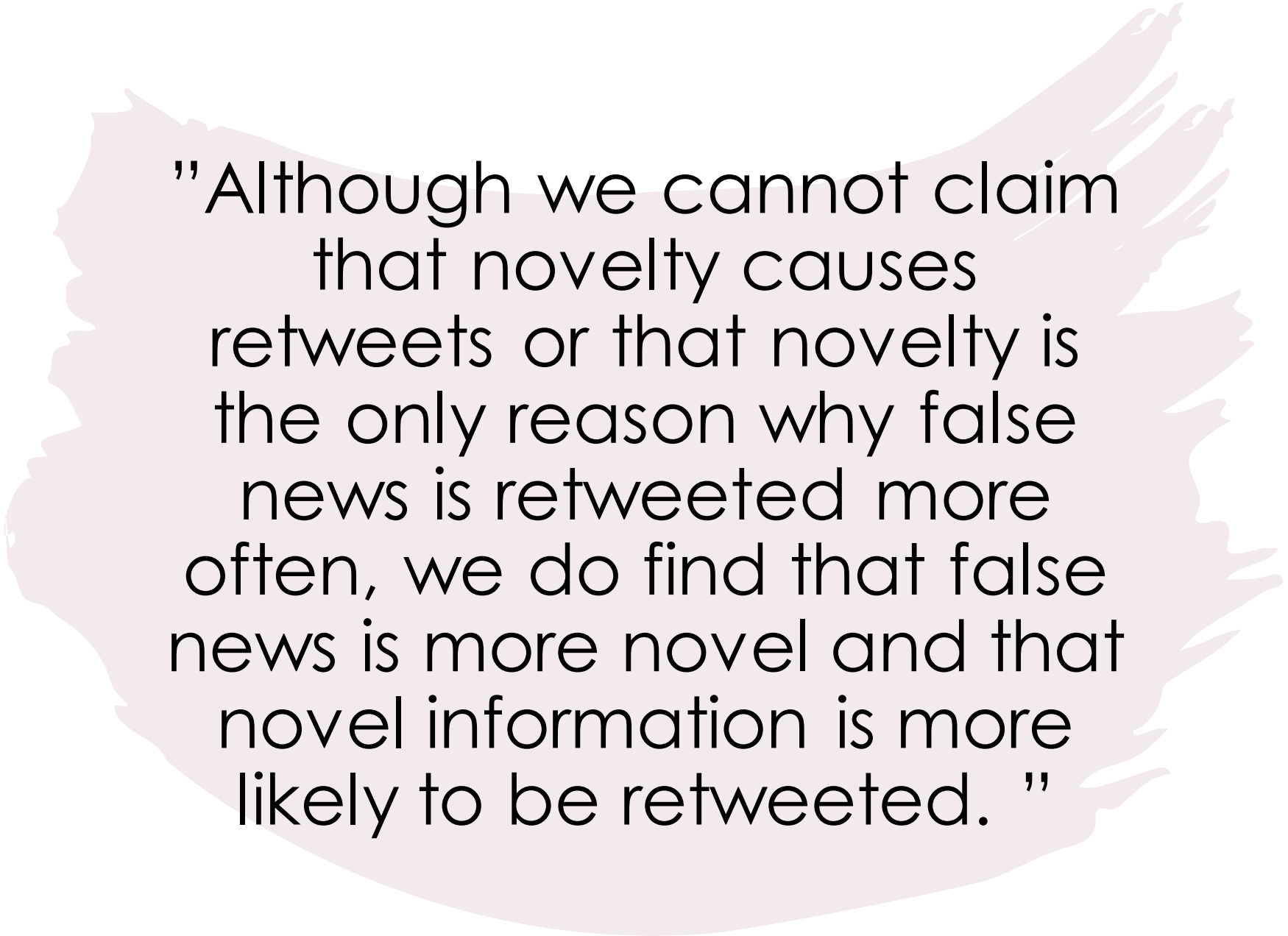
- Users' perceptual emotions?

Assess users' perceptions of the information contained in true and false rumors by comparing the emotional content of replies to true and false rumors.

False rumors inspired replies expressing greater surprise corroborating the novelty hypothesis, and greater disgust, whereas the truth inspired replies that expressed greater sadness, anticipation, joy and trust.

	mean		variance		ks-test
	false—true	false—true	false—true	false—true	
surprise	0.172	0.116	0.0167	0.0072	D=0.205, p~0.0
disgust	0.240	0.205	0.0260	0.0227	D=0.102, p~0.0
fear	0.108	0.102	0.0120	0.0095	D=0.021, p~0.164
anger	0.122	0.126	0.0074	0.0111	D=0.023, p~0.078
sadness	0.061	0.068	0.0038	0.0065	D=0.037, p~0.0
anticipation	0.140	0.150	0.0093	0.0154	D=0.038, p~0.0
joy	0.071	0.087	0.0054	0.0104	D=0.061, p~0.0
trust	0.087	0.104	0.0058	0.0119	D=0.060, p~0.0

The emotion in the replies is categorized by using the leading lexicon, which provides a comprehensive list of ~140,000 English words and their associations with eight emotions above, and a list of ~32,000 Twitter hashtags and their weighted associations with the same emotions.



”Although we cannot claim that novelty causes retweets or that novelty is the only reason why false news is retweeted more often, we do find that false news is more novel and that novel information is more likely to be retweeted. ”

The background of the slide is a dense collection of small, colorful wooden human figures. The figures are in various colors including light wood, orange, red, dark wood, and teal. They are scattered across the entire frame, creating a textured and vibrant background. A white, torn-edge paper shape is overlaid in the center, containing the text.

Methods & Analyses – Robustness Test

- Robustness to the cascade clustering errors
- Robustness to selection bias of the news source
- Robustness to the exclusion of bots

Methods & Analyses – Robustness Test

One Rumor <- cascades belonging to the same rumor are clustered together

errors in clustering cascades

Robustness to the cascade clustering errors

By comparing analyses with and without the clustered errors, they found that, although clustering reduced the precision of our estimates as expected, the directions, magnitudes, and significance of their results did not change.

Methods & Analyses – Robustness Test

Robustness to selection bias of the news source

Rumors <- selected from the six fact-checking organizations

So they **independently verified a second sample of rumor cascades** that were not verified by any fact-checking organization. These rumors were fact checked by three undergraduate students.

It is found that the results are nearly identical to those estimated with the main dataset.

Methods & Analyses – Robustness Test

Robustness to the exclusion of bots

Rumors <- some are made or spread by bots

They used a bot-detection algorithm to identify and remove all bots before running the analysis. They found that none of the main conclusions changed.

Although the inclusion of bots accelerated the spread of both true and false news, it affected their spread roughly equally. This suggests that false news spreads farther, faster, deeper, and more broadly than the truth because humans, not robots, are more likely to spread it.



I would strongly recommend this paper because:

- The paper confirms that false news spreads more pervasively than the truth online. It also overturns conventional wisdom about how false news spreads.



Q&As

The background of the slide is a dense collection of small, colorful wooden human figures. The figures are in various colors including light wood, orange, red, dark wood, and teal. They are scattered across the entire frame, creating a textured and vibrant background. A white, torn-edge paper graphic is overlaid in the center, containing the text.

Discussions

- Q1: How do you feel about news that is partially true and news that cannot necessarily be fact-checked?
- Q2: What do you think is more important? The truth of something, or the majority's belief about it?
- Q3: Do you have any experience or observations of others spreading false news online? Tell us more about it.

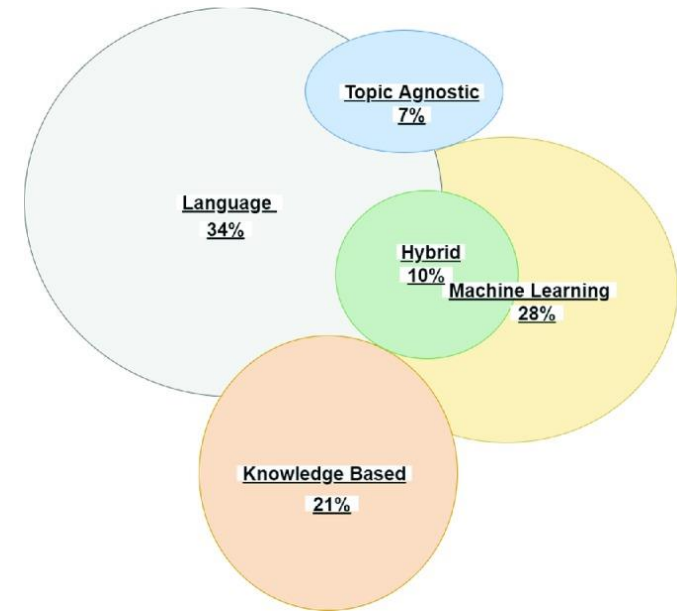
The Limitations of Stylometry for Detecting Machine- Generated Fake News

TAL SCHUSTER, ROEI SCHUSTER, DARSH J.
SHAH, REGINA BARZILAY



Why is this topic important?

- Stylometry – extraction of stylistic features from written text
 - Detects provenance of text to prevent impersonations
 - Detects misinformation due to deception
- Human-impersonating neural language models (LMs) can mass-produce both malicious and helpful text
 - Malicious: misinformation through impersonation/fallacious or misleading
 - Helpful: text auto-completion, auto question answering
- Stylometry-based approaches have proven to defend malicious human-written text
 - However, not much study has been done on using stylometry to distinguish malicious LMs from legitimate one.



The Overview of the Paper

- Generate Data set
 - Define "fake news"
 - Collect labeled data
 - Extension Dataset
 - Modifications Dataset
- Train and Evaluate the model
 - Grover-Mega discriminator (ML-based)
- Result Analysis
 - Fail to detect Misinformation
 - Detect Human-Machine Impersonation

Methods

Data generation methods

- Extension dataset

Generated by Grover's generator

newsQA dataset

CNN article

Corresponding
Questions and
Answers

(a) QA extension

Title: Fernandez defends Argentine grain export tax

President Cristina Fernandez on Tuesday defended an increase in export taxes on grains that has riled many farmers, and she called on them to respect the law in protesting her policies.

<...> In a concession to her critics, Fernandez said the increase in taxes on exports of grains that she instituted in March by decree will be debated by Congress. But there is little likelihood that the Congress will order major changes, since her party controls both houses.

But Hilda Duhalde, an opponent of Fernandez, was not persuaded. "It's true that they have a majority in both houses, but we have to put white on black and watch out for the small- and medium-sized producers, who are the ones suffering," she said.

Argentina raised export taxes in March by more than 10 percent. Fernandez has said growers have benefited from rising world prices and the profits should be spread to help the poor.

Farmers have countered that they need to reinvest the profits and that the higher taxes make it difficult for them to make a living.

Fernandez said she was open to dialogue, but a dialogue that does not countenance the blocking of roads or other disruptions to the lives of Argentines. "Democracy for the people, not the corporations," she said.

We attempt to answer: Who appealed for dialogue and respect?

Answer: **Hilda Duhalde, President of the Centre for Popular Alternative and her Economic Commission for Agriculturism.** *(fake; President Cristina Fernandez)*

We attempt to answer: What do farmers say higher taxes do? **Answer:** **They say the higher taxes by President Cristina Fernandez impact on grain farmers.** *(real)*

Methods

Data generation methods - Modification dataset

GPT-2 Medium LM* on NYT articles

Random deletion
and insertion

($m = 2, 6, 10$ used)

- $m/2$ negation deleted
- $m/2$ negation inserted

Preserve total number
of negations

Labeling

- Original: "True"
- Modified: "Fake"

(b) Article modification

Title: Nominee Betsy DeVos's Knowledge of Education Basics Is Open to Criticism

Until Tuesday, the fight over Betsy DeVos's nomination to be secretary of education revolved mostly around her support of contentious school choice programs. But her confirmation hearing that night opened her up to new criticism: <...> Ms. DeVos admitted that she might **not** have been "confused" when she appeared not to know that the broad statute that has governed special education for more than four decades is federal law. <...> She appeared blank on basic education term: Asked how school performance should be assessed, she **did not** know the difference between growth, which measures how much students have learned over a given period, and proficiency, which measures how many students reach a targeted score. Ms. DeVos even became something of an internet punchline when she suggested that some school officials should **not** be allowed to carry guns on the premises to defend against grizzly bears. <...> But her statements on special education could make her vulnerable families of children with special needs are a vocal lobby, one that Republicans **do not** want to alienate. <...> Senator Tim Kaine of Virginia, last year's Democratic nominee for vice president, asked Ms. DeVos whether schools that receive tax dollars should be required to meet the requirements of IDEA. "I think that is a matter that's best left to the states," Ms. DeVos replied. Mr. Kaine came back: "So some states might be good to kids with disabilities, and other states might not be so good, and then what? People can just move around the country if they don't like how their kids are being treated?" Ms. DeVos repeated, "I think that is an issue that's best left to the states." "It's **not** federal law," an exasperated Mr. Kaine replied. <...> "Do you think families should have recourse in the courts if schools don't meet their needs?" she asked. "Senator, I assure you that if confirmed I will be very sensitive to the needs of special needs students," Ms. DeVos said. "It's **not** about sensitivity, although that helps," Ms. Hassan countered. <...>

GPT-2 Language Model

- **A large-scale unsupervised language model**
 - Simply trained to predict the next word
- **Generates coherent paragraphs of text**
 - Without domain-specific training datasets
- **Also able to perform...**
 - Rudimentary reading comprehension
 - Machine translation
 - Question answering
 - Summarization



Machine Learning

Methods

Data generation methods

- Automatic Article Extension (vanilla) dataset

GPT-2 Medium LM on NYT articles

Conditioned on first 500 words from NYT article (label: "real text")

Automatically Extended w.r.t **g** (label: "fake text")

g: percentage of machine-generated text

(c) Vanilla extension

SEOUL, South Korea — North Korea's leader, Kim said on Sunday that his country was making final preparations to conduct its first test of an intercontinental ballistic missile — a bold statement less than a month before the inauguration of Donald J. Trump. Although North Korea has conducted five nuclear tests in the last decade and more than 20 ballistic missile tests in 2016 alone, and although it habitually threatens to attack the United States with nuclear weapons, the country has never an intercontinental ballistic missile, or ICBM. <...> In his speech, Mr. Kim did not comment on Mr. Trump's election. Doubt still runs deep that North Korea has mastered all the technology needed to build a reliable ICBM. But analysts in the region said the North's launchings of rockets to put satellites into orbit in recent years showed that the country had cleared some key technological hurdles. After the North's satellite launch in February, South Korean defense officials said the Unha rocket used in the launch, if successfully reconfigured as a missile, could fly more than 7,400 miles with a warhead of 1,100 to 1,300 pounds — far enough to reach most of the United States. **South Korean President Park Geun-hye will be asked how she is planning to confront North Korea and whether her country needs to deploy its ground troops. It also is unlikely that she will deploy U.S. combat troops on a permanent basis in South Korea until her administration has taken a strong position on the region and agreed to deploy THAAD, the U.S. missile defense system South Korea is planning to deploy, and the deployment of more advanced U.S. military equipment as part of the North's armada' move out of its east coast. Mr. Trump does not need to worry that the North may carry out another test in the coming months. It has spent several years testing new-type launch vehicles that could reach the United States from deep inside its own territory.**

Methods

Evaluation methods – Detecting Machine- Generated Misinformation

- Generated answers were manually labeled real or fake by correctness with nonsensical ones (29%) filtered out
- Two different annotators were runThe labels from two different annotators were compared (inter-annotator agreement) and substantial (Cohen's kappa score of $k=0.78$)
- Removed highest TF-IDF-weighted word-count similarity containing the answer from each article

Methods

Evaluation methods – Performance Metrics

- Classification Performance evaluation using Confusion Matrix
- Use of precision, recall, F1, accuracy scores to evaluate the detector's performance against both data sets and human control
- TP – fake news is detected as fake
- FN – fake news detected as real
- FP – real news detected as fake
- TN – real news detected as real

		True condition	
		Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive	False positive, Type I error
	Predicted condition negative	False negative, Type II error	True negative

$$\begin{aligned} \textit{precision} &= \frac{TP}{TP + FP} \\ \textit{recall} &= \frac{TP}{TP + FN} \\ \textit{F1} &= \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \\ \textit{accuracy} &= \frac{TP + TN}{TP + FN + TN + FP} \end{aligned}$$

Methods

Evaluation methods –
Detecting Machine-Human
Impersonations

- Zero-shot Setting
 - Use of zero-shot classifier to detect fully machine generated articles
 - Applied on full article, vanilla extension (g=20% & g=1%)
- Adaptive Setting
 - Grover detection is measured on article extension generations to see how effective detection of human from machine is
 - Applied on full article, vanilla extension (g=20% & g=1%), and QA extension

Analyses

- Result (Section 4)
 - Stylometry fails to detect Machine-generated misinformation
 - F1 score of extension and modification datasets are better than majority baseline of 51%.
 - However, does not perform much better than humans in detecting potential misinformation and that if humans verify against other resources, it will drastically improve results (F1: 0.69 → 0.84).

Machine-generated misinformation		precision	recall	F1	accuracy
adaptive	QA extension (false vs. true)	0.72	0.71	0.71	71%
	modification ($m = 2$)	0.53	0.52	0.53	53%
	modification ($m = 6$)	0.66	0.65	0.65	65%
	modification ($m = 10$)	0.73	0.47	0.63	65%

Table 1: Results: Section 4. We report (macro) F1 score and overall accuracy, as well as precision and recall of the “fake” class. Zero-shot performance (not included) was very low in all cases.

QA Extension on Humans	F1 Score
First Subject	0.68
Second Subject	0.84

Analyses

- Result (Section 5)
 - Stylometry Detects Machine-Human Impersonations
 - Effectively distinguishes human from machine
 - For full article, both zero-shot and adaptive setting reach 0.9 or higher F1 score

Provenance detection		precision	recall	F1	accuracy
zero-shot	full article	0.84	0.98	0.90	90%
	vanilla extension ($g = 20\%$)	0.52	0.20	0.45	51%
	vanilla extension ($g = 1\%$)	0.07	0.01	0.28	37%
adaptive	full article	0.93	0.94	0.94	94%
	vanilla extension ($g = 20\%$)	0.90	0.97	0.95	95%
	vanilla extension ($g = 1\%$)	0.91	0.95	0.94	95%
	QA extension (machine vs. human)	0.82	0.86	0.83	83%

Analyses

- Result (Section 5)
 - Stylometry Detects Machine-Human Impersonations
 - Effective in preventing impersonation, but has limited scope
 - Might not reflect the true performance on versatile LM
 - LM only used for generating fake news in this paper
 - F1 score drops for adaptive setting on QA extension due to the criterion by which the template for QA was selected, human "reasonableness" score

Provenance detection		precision	recall	F1	accuracy
zero-shot	full article	0.84	0.98	0.90	90%
	vanilla extension ($g = 20\%$)	0.52	0.20	0.45	51%
	vanilla extension ($g = 1\%$)	0.07	0.01	0.28	37%
adaptive	full article	0.93	0.94	0.94	94%
	vanilla extension ($g = 20\%$)	0.90	0.97	0.95	95%
	vanilla extension ($g = 1\%$)	0.91	0.95	0.94	95%
	QA extension (machine vs. human)	0.82	0.86	0.83	83%

Strength of the Paper

- Generation of the two benchmarks
 - Two different criteria based on two common applications of stylometry:
 - Detecting the **provenance** of text to prevent impersonations
 - Detecting **misinformation** in text due to deception
- Point out that stylometric approach is not completely sufficient
 - Effective in preventing impersonations but shows limited performance in detecting LM-generated misinformation
 - Unable to detect stylistic differences between fallacious and genuine content when LMs are used to generate both
- Motivates future research on:
 - Constructing more benchmarks for NLP-based approaches
 - Improving non-stylistic methods
 - Interdisciplinary field beyond NLP

Weakness of the Paper

- The misinformation evaluated in the paper may not accurately represent the misinformation that exists in real life social media
- More metrics within derived from the confusion matrix could be calculated and shown for improved insights on performance
- Reader does not have much information on people selected to perform “human evaluation” – could be a source of selection bias

The background of the slide is a dense collection of small, colorful wooden human figures. The figures are in various colors including light wood, orange, red, dark brown, and teal. They are scattered across the entire frame, creating a textured and vibrant background. A white, torn-edge paper overlay is positioned in the center, containing the text.

Discussions

- Q1: Do you know of any other methods of fake news detection? If so, what is the context?
- Q2: What are your experiences with machine generated fake news? Were you able to identify whether it was a machine? If so, how what features allowed you to do it?



Q&As



Thank you!