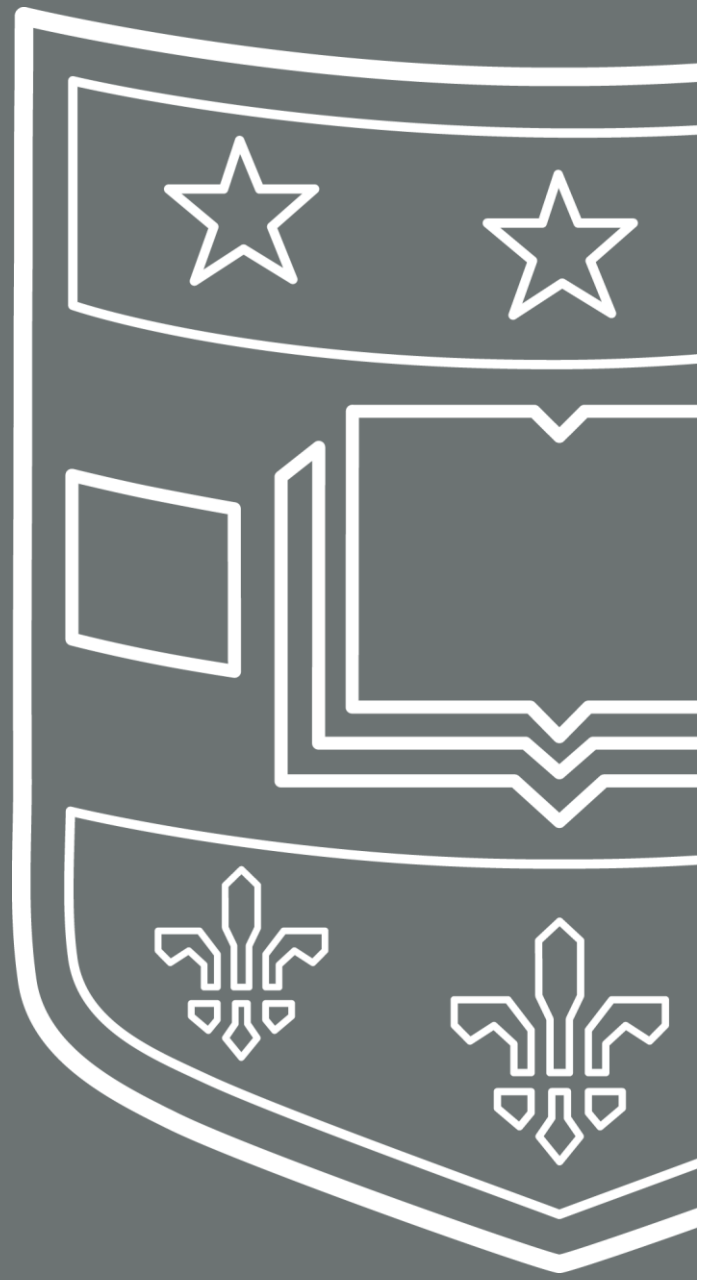
Twitter's Birdwatch: A case study of ~~the impact of~~ crowd-sourced and community-based fact-checking ~~on the spread of misinformation~~

Merriah Croston, MPH
CIRN Conference Prato
November 10, 2022

Washington University in St. Louis



Preliminaries

What is Twitter? What is Birdwatch?

User

Tweet

Engagement

BW Note
+
BW Ratings

Note

Initial study aim

Examine the impact of Birdwatch on the spread of misinformed tweets by comparing how misinformed tweets spread before and after flagging

Unfortunately…

Tweets are flagged via the Birdwatch mechanism **2 hours to 60 days** after they are posted; however, the average tweet stops spreading in **less than 30 minutes.**

Revised study aim

Examine the output of Birdwatch by comparing misinformed tweets that are flagged to those that are not flagged

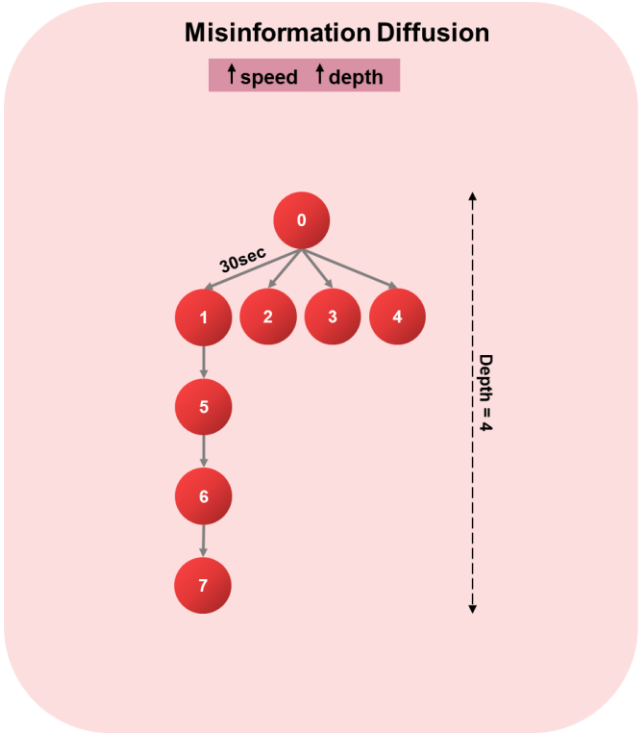
Research questions

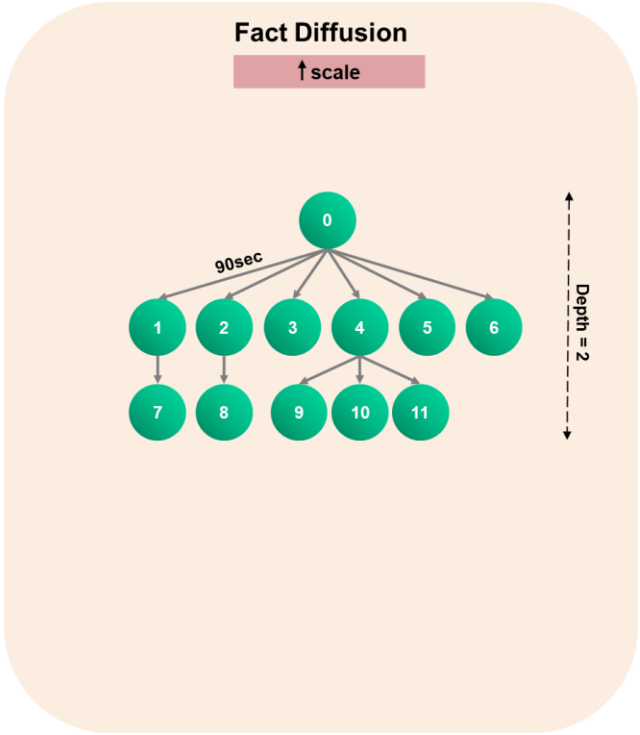
RQ1: How do misinformed tweets that have been flagged with a note differ from those that are not flagged?

RQ2: How do misinformed tweets that are flagged early differ from those that are flagged late?

Misinformation on social media

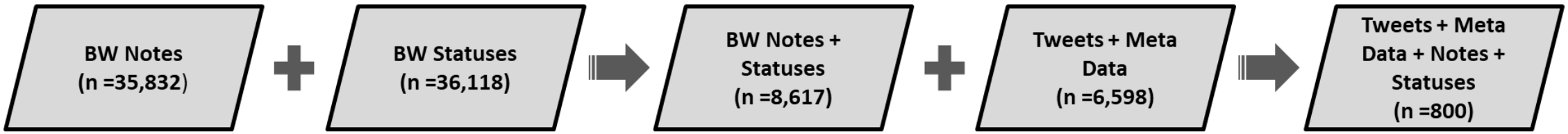
- Misinformation spreads faster and deeper than factual information. [1]
- Misinformation affects cognitions, affect, behavior, and health outcomes. [2]
- Fact-checking has a modest effect on how misinformation spreads and the outcome of receiving misinformation. [3,4]





Methods & Results

Birdwatch + Twitter archive data



Birdwatch + Twitter archive data

Birdwatch Data
Tweet flagged
Tweet flagged early
Tweet potential harm (little vs. considerable)
Tweet misinformation type: Manipulated media Outdated information Missing important context Unverified claims
Time from BW note creation to data retrieval (days)

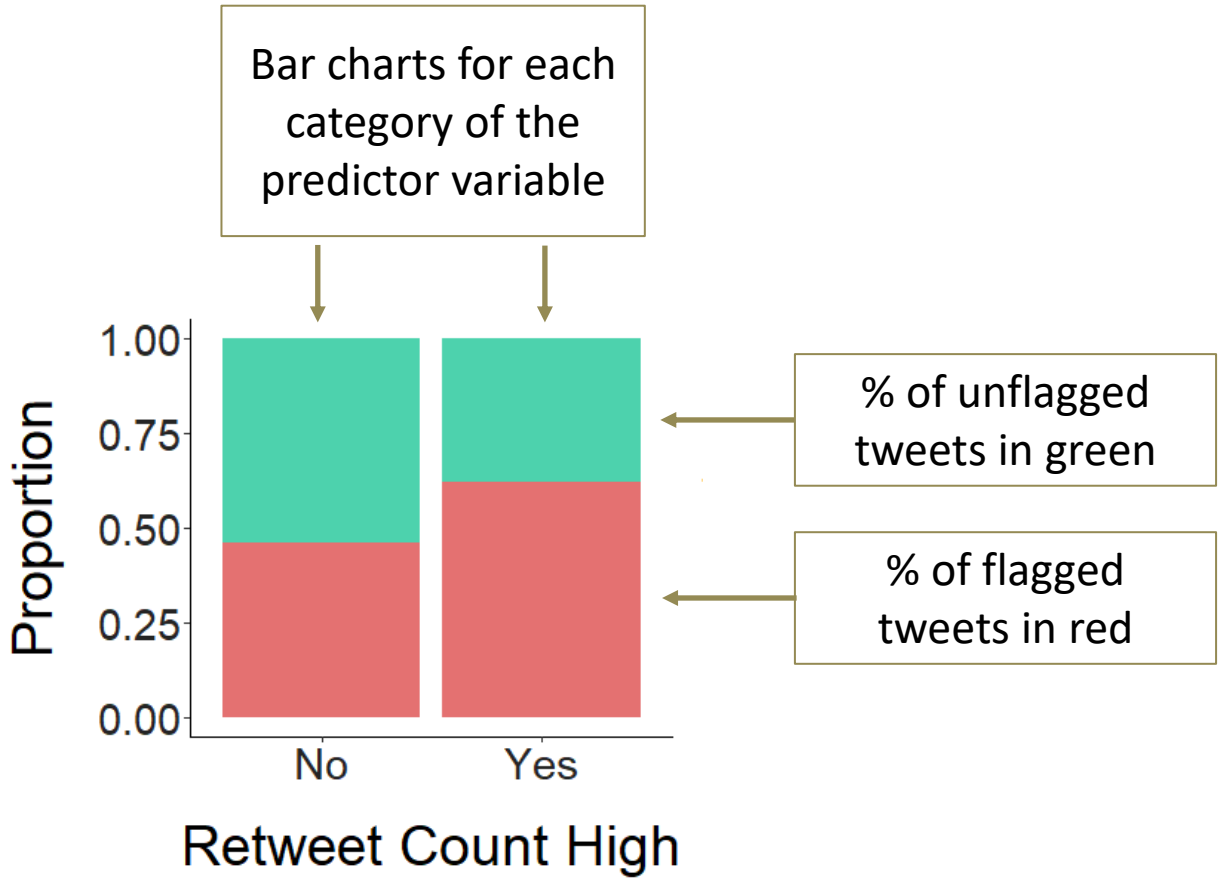
+

Twitter Archival Data
User characteristics: Follower count Tweet count Account age (years)
Tweet characteristics: High retweet count High quote count High like count

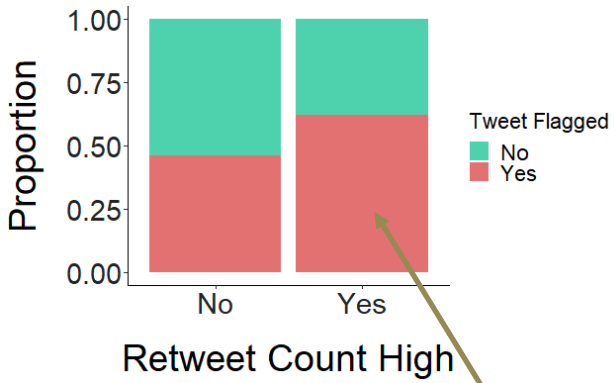
Research question 1:

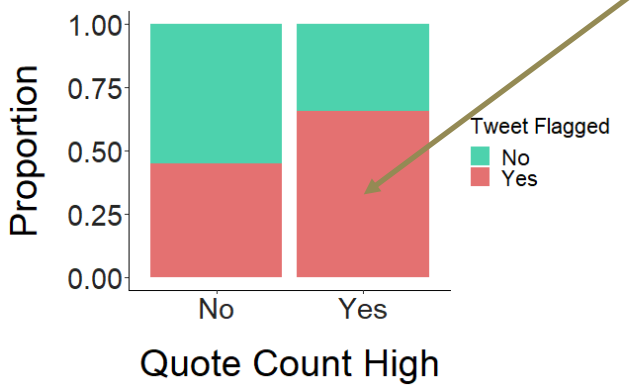
How do misinformed tweets that have been flagged with a note differ from those that are not flagged?

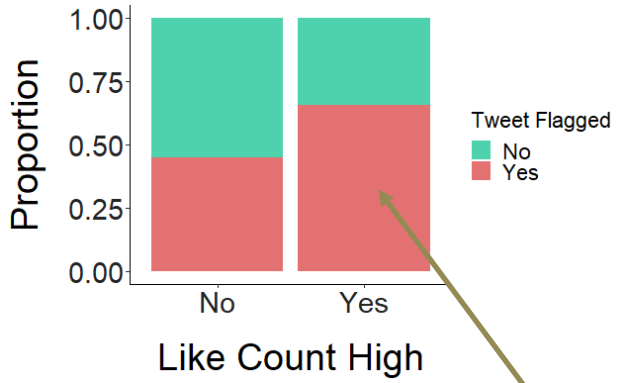
Examining relationships between tweet flag status and categorical predictor variables



High retweet count, high quote count, and high like count are significantly associated w/ a tweet's flag status







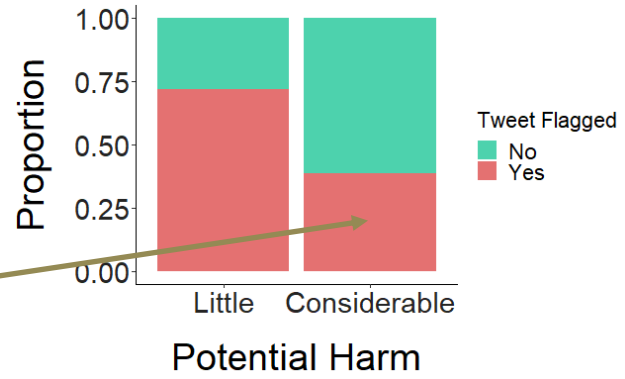
High quote count higher odds of being flagged than low quote count

High retweet count higher odds of being flagged than low retweet count

High like count higher odds of being flagged than low like count

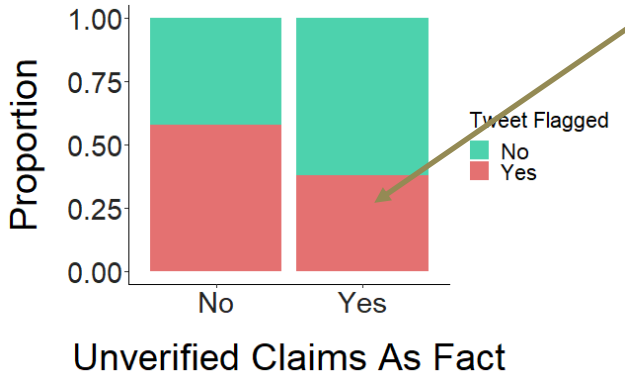
Potential harm and a few misinformation types are significantly associated w/ a tweet's flag status

Tweets w/ considerable potential harm lower odds of being flagged



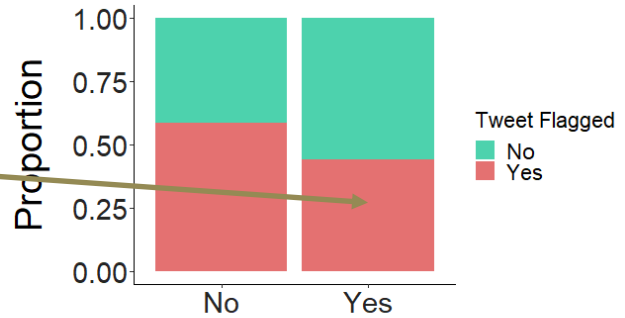
Potential Harm

Tweets w/ unverified claims lower odds of being flagged



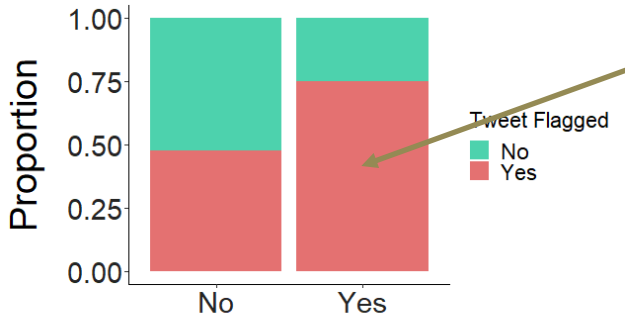
Unverified Claims As Fact

Tweets missing important context lower odds of being flagged



Missing Important Context

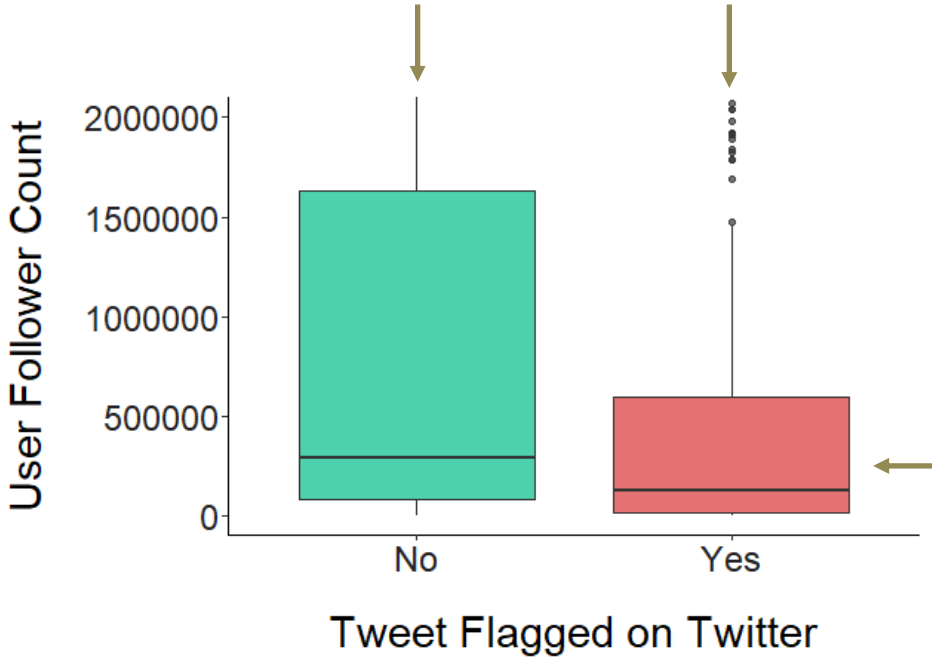
Tweets w/ manipulated media higher odds of being flagged



Includes Manipulated Media

Examining relationships between tweet flag status and continuous predictor variables

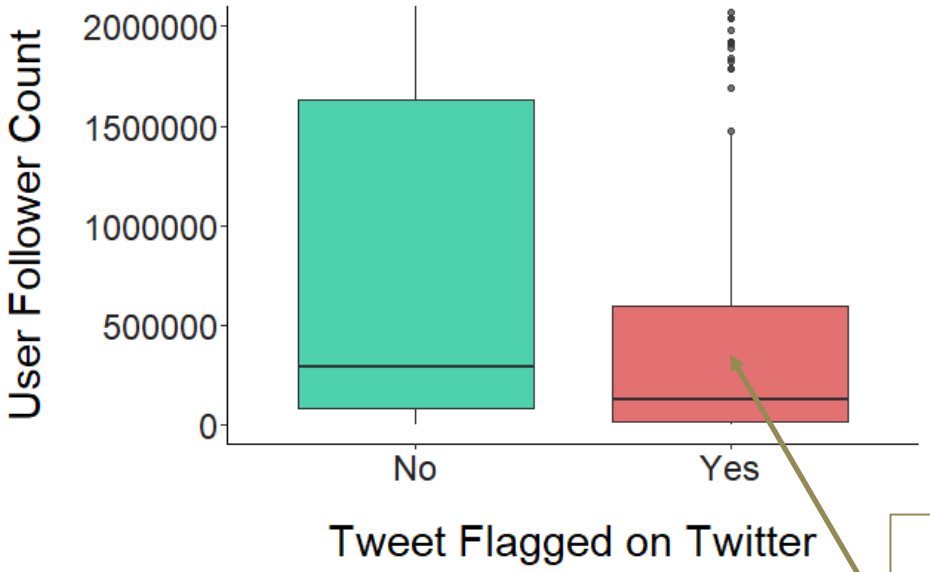
Box-whisker plots are shown for flagged tweets on the right and unflagged tweets on the left.



The colored box in each plot represents the values for 50% of tweets.

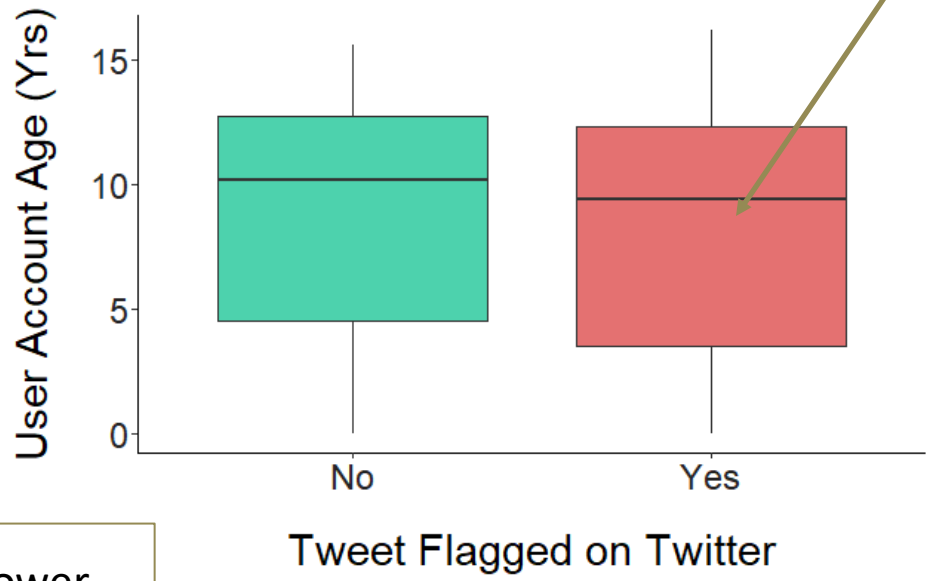
The horizontal line in each box represents the median value of the predictor variable.

User follower count and account age are significantly associated w/ a tweet's flag status

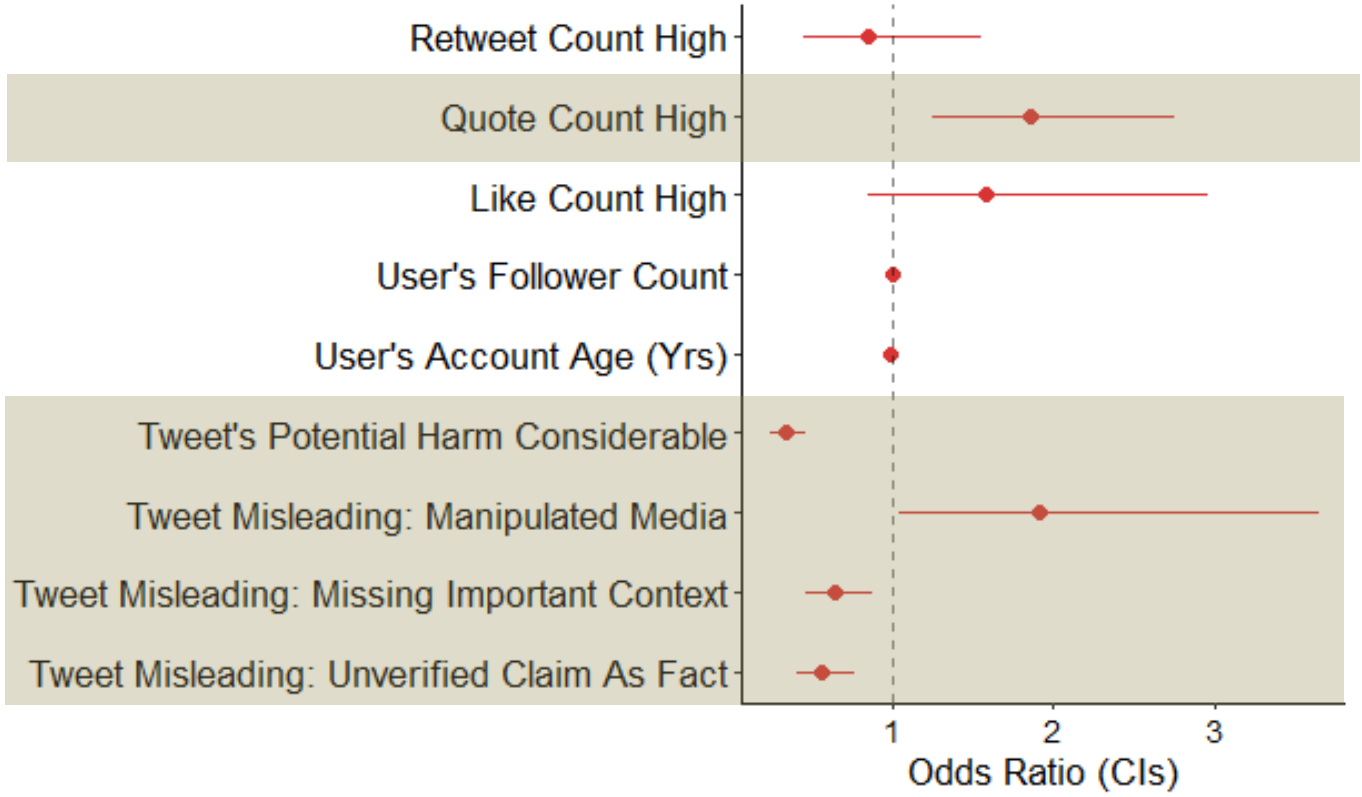


Lower follower counts higher odds of being flagged

Younger account age higher odds of being flagged



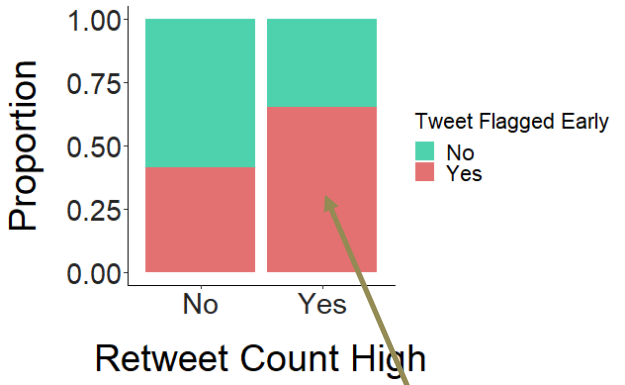
Quote count, harm & misinformation type variables still significant associated w/ flag status when controlling for other variables



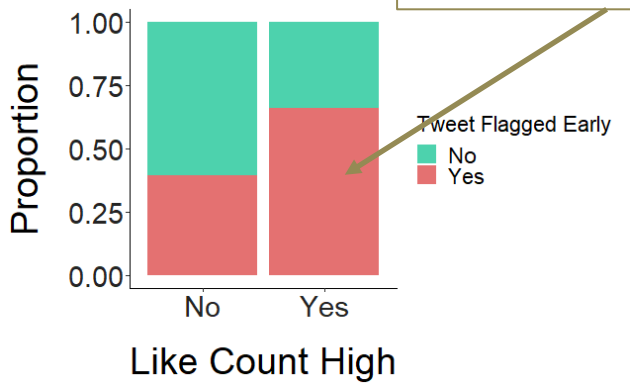
Research question 2:

How do misinformed tweets that are flagged early differ from those that are flagged late?

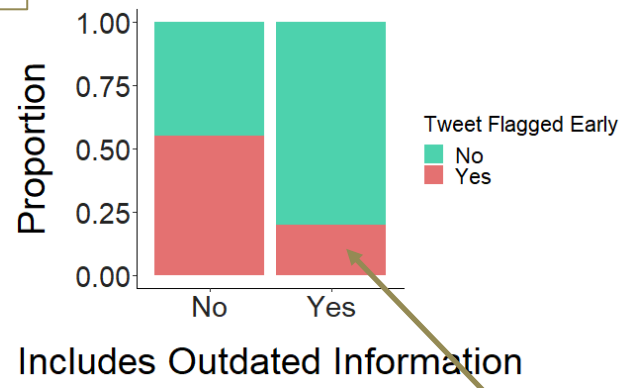
High retweet count, high quote count & one misinformation type significantly associated w/ a tweet's early flag status



High retweet count higher odds of being flagged early

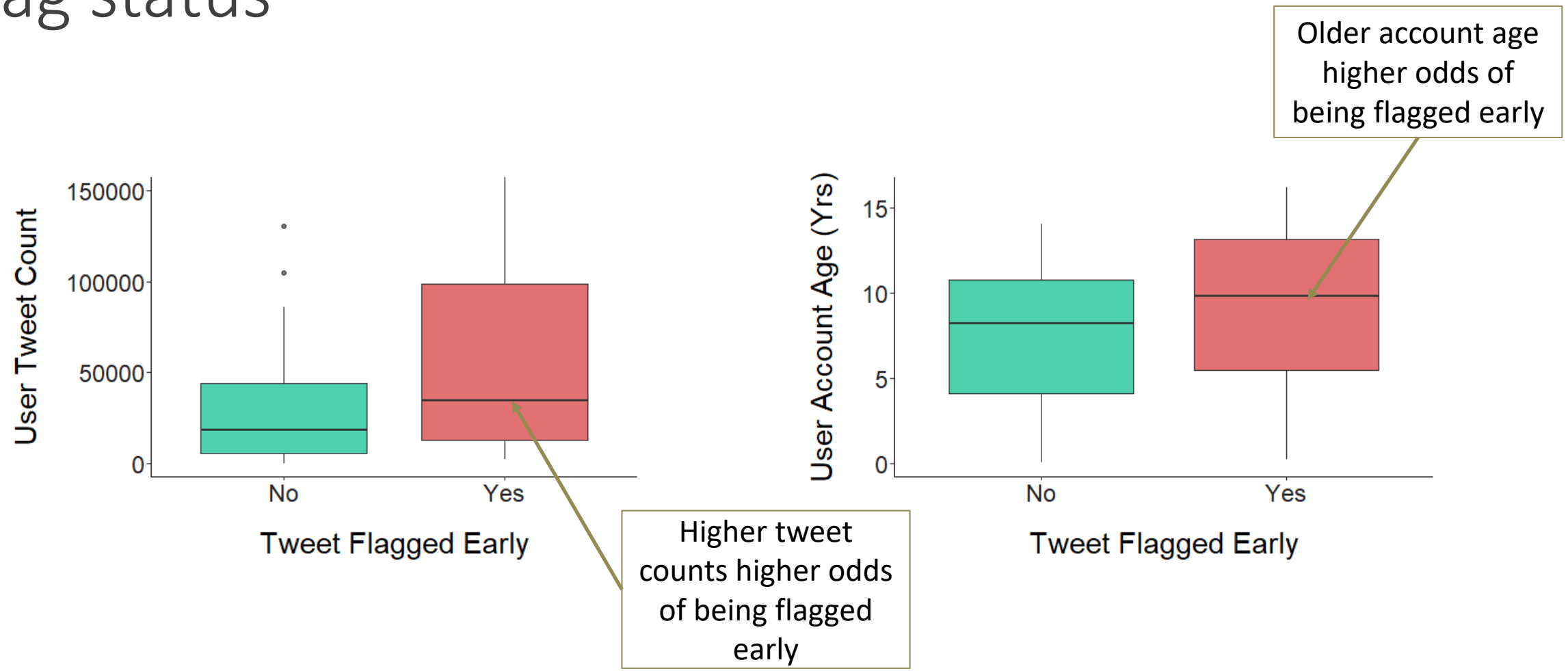


High like count higher odds of being flagged early



Tweets w/ outdated info lower odds of being flagged early

User tweet count and account age are significantly associated w/ a tweet's early flag status



Discussion

Conclusions & Implications

- High quote count & misinformation type most robust predictors of flagging
- Surprisingly, tweets w/ considerable potential harm have lower odds of being flagged; tweets missing important context or w/ unverified claims have lower odds of being flagged, while tweets w/ manipulated media have higher odds of being flagged.
  - Does this indicate biases in the BW rating process?
  - Does this indicate that Birdwatchers have more difficulty rating or reaching consensus on certain misinformation categories or considerably harmful misinformation?
- Similar, but distinct, predictors of flagging and early flagging
  - Perhaps, unflagged tweets are not misinformed. If true, I compared misinformed tweets to verifiable tweets in pursuit of the first research question.

What's next?

- Clarifying the questions/speculations in the last slide
  - Does this indicate biases in the BW rating process?
  - Does this indicate that Birdwatchers have more difficulty rating or reaching consensus on certain misinformation categories or considerably harmful misinformation?
  - Are unflagged tweets misinformed, despite Birdwatchers' original designations?
- Investigate additional predictors of flagged and flagged early
  - Characteristics of the Birdwatcher who reported the tweet
  - Semantic predictors
  - Bot analysis
- If more data become available and if the timing of flagging improves
  - Multivariable analysis of early flagging
  - Social network analysis of the impact of flagging and early flagging on misinformation diffusion

Literature

1. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151. http://science.sciencemag.org/
2. Romer, D., & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Social Science and Medicine*, *263*. https://doi.org/10.1016/j.socscimed.2020.113356
3. Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, *36*(13), 1776–1784. https://doi.org/10.1080/10410236.2020.1794553
4. Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Thank you!

Merriah Croston, MPH
mcroston@wustl.edu
@Merriah_WithAnE