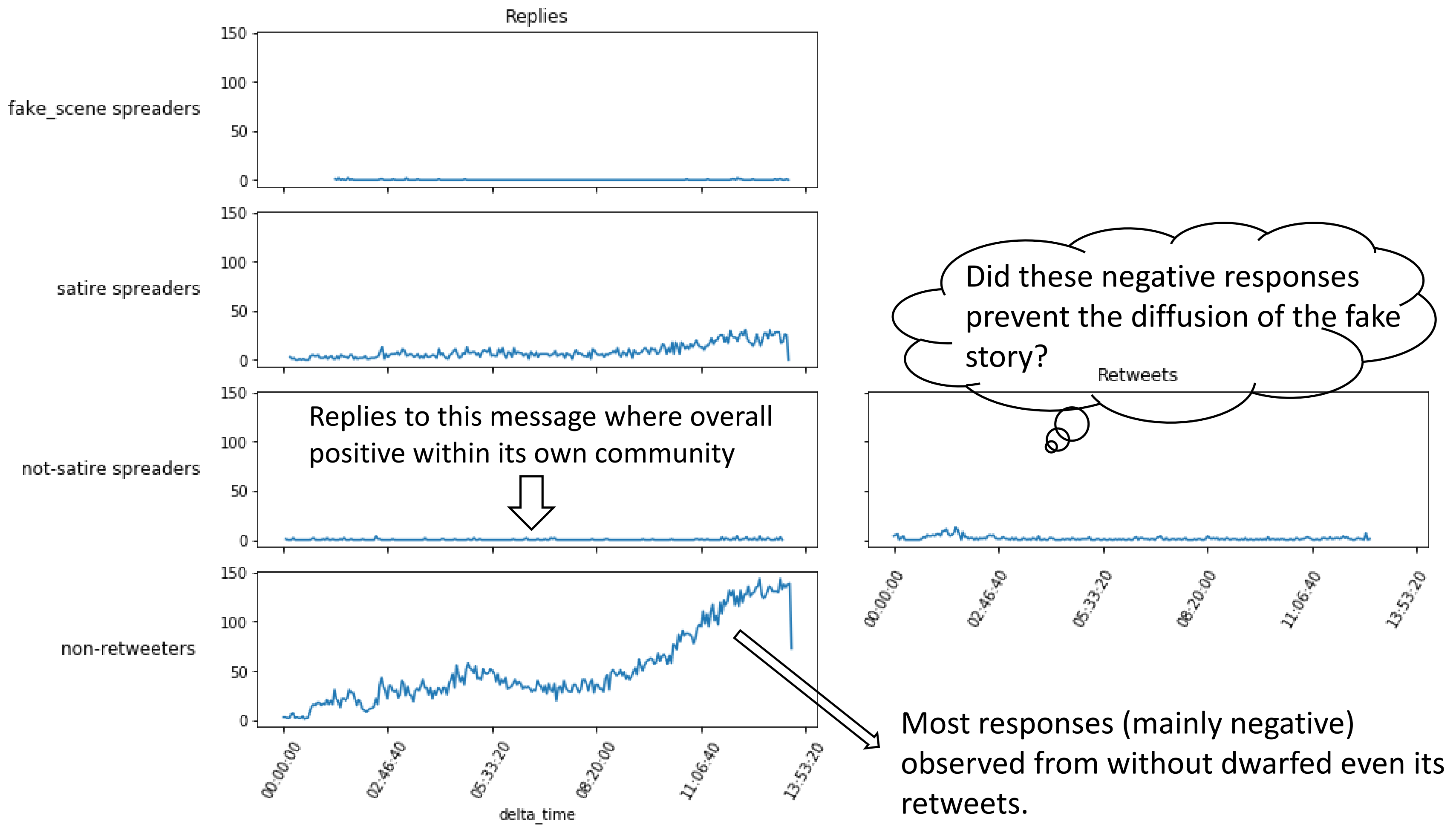
Ramon Alfonso Villa-Cox
rvillaco@andrew.cmu.edu

Carnegie Mellon University
www.casos.cs.cmu.edu

TwitterSim: A policy-oriented test-bed for the spread of Contentious Messages in Twitter



Replies to this message where overall positive within its own community

Did these negative responses prevent the diffusion of the fake story?

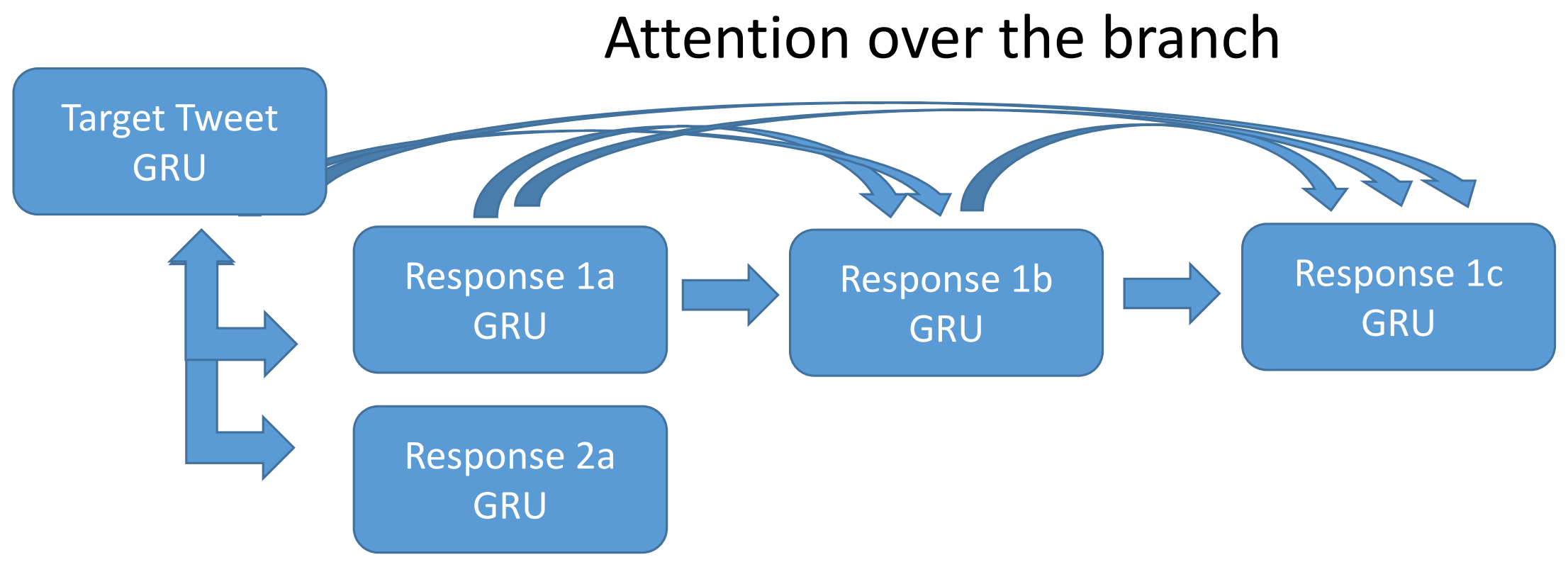
Most responses (mainly negative) observed from without dwarfed even its retweets.

- Black Panther opened to much social media fanfare and financial success in Feb 2018.
- The film was hyped on social media due in part to its representation of African and African-American actors and creators.
- We identified a total of four types of false stories that were shared and reacted to on Twitter.
- User communities were defined based in the retweeters of each type of story (as it signals tacit approval of the message).

Type of False Story	Description
Fake Attacks (Non-Satire)	claimed race-based assaults at movie theaters used images that were debunked by community
Fake Attacks (Satire)	mocked the original fake attack posts used more unbelievable images from pop-culture
Fake Scene	claimed movie contained false sexual/racial scenes
Alt-Right	claimed the movie promoted alt-right philosophy

Identifying the Stance of Responses in Twitter

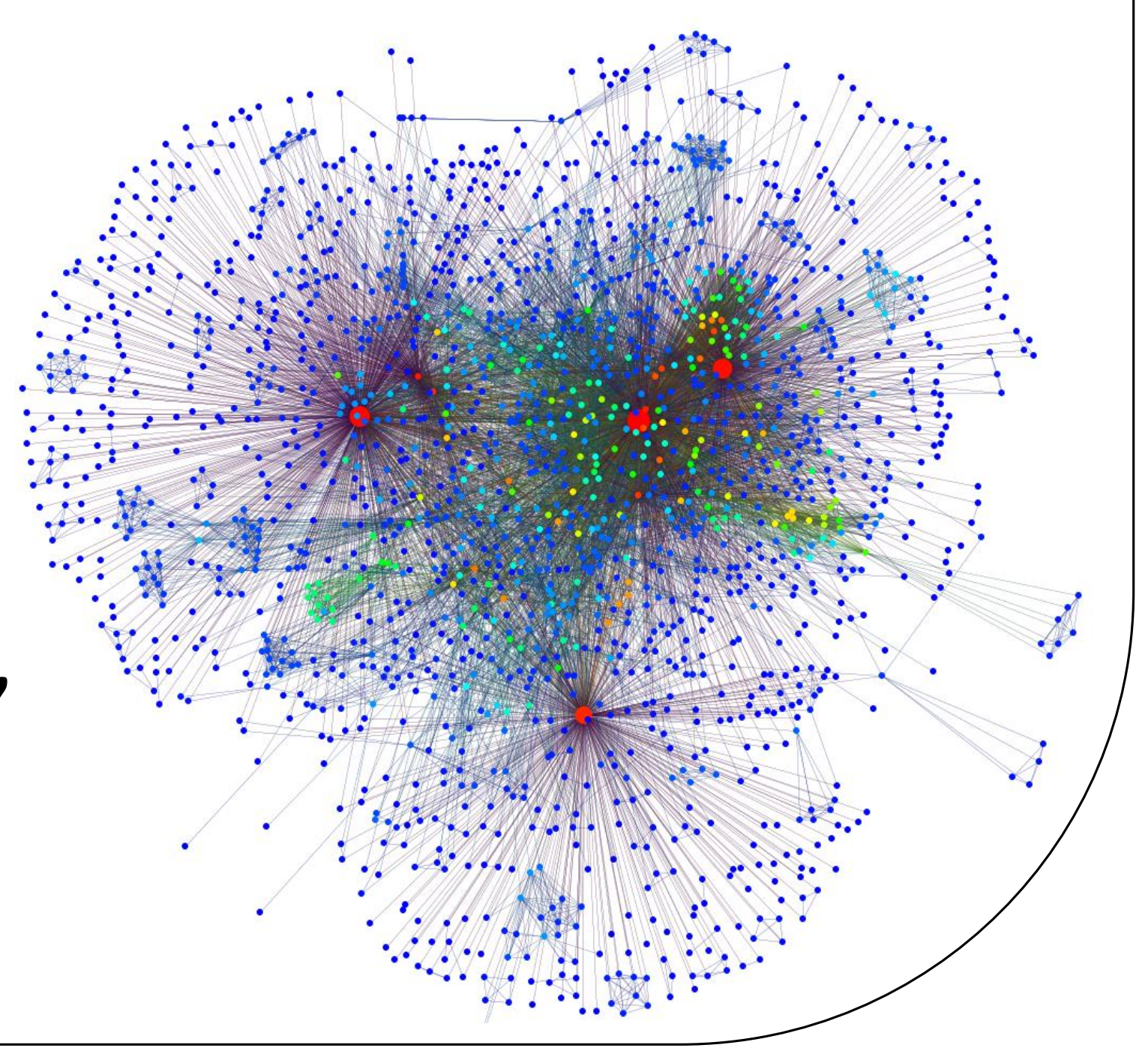
- We developed a neural network classifier that uses the conversational structure of Twitter threads to classify the stance of responses.
  - Responses are classified as commenting, supporting, denying or querying.


Attention over the branch

- We seek to improve on the state of the art in two ways, by including an attention mechanism over conversation threads.
  - This can improve accuracy over longer threads.
- We also designed a collection methodology oriented towards denials (which is considerably under-sampled in available datasets).

TwitterSim

- A simulation that incorporates the structural properties of the platform as well as mechanisms for individual user behavior.
- These properties impose a series of rules for interaction and content promotion that influence the diffusion of information.
- We model the main types of interactions available:
  - Quotes, Replies and Retweets
- An economy of attention is introduced by modelling the timeline of each users and their limited capacity to read them.
- The behavioral model of agents allow for stance in the type of responses in order to explore the effect of these interactions on the diffusion of contentious information.



- **Validation:** we will evaluate the simulated diffusion process of the different types of stories against what is observed on the different case studies.
- The case studies, based on known rumors, are determined based on our stance classifier.

This work was supported in part by the Office of Naval Research (ONR) through grants N000141812108 and N00014-17-1-2605 / 25323818CMU and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the ONR or the U.S. government.