The Structure of Toxic Conversations on Twitter

Martin Saveski, Brandon Roy, Deb Roy
Massachusetts Institute of Technology

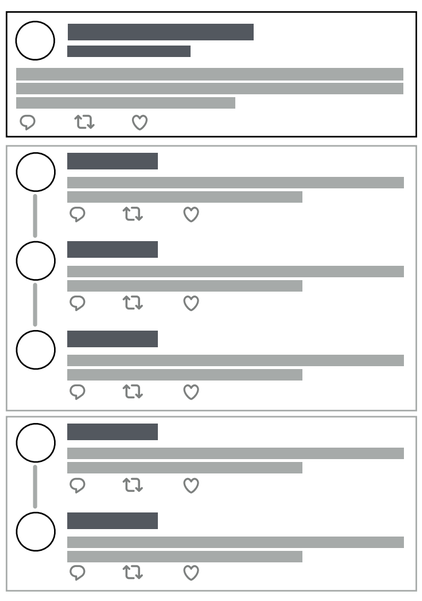
Goals

- Analysis: study the relationship between structure and toxicity of conversations, after the conversations are over
- Prediction: predict future toxicity based on the structure of the conversation, as the conversation unfolds

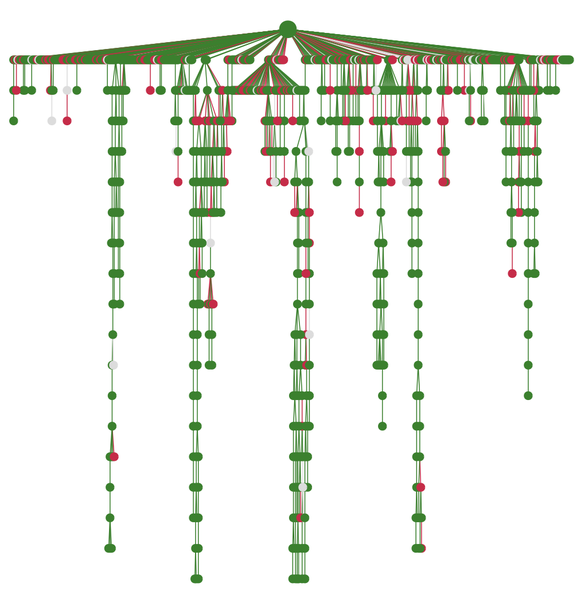
Data

- News: 510K+ conversations, 32M+ tweets, 5 outlets, 1 year
- Midterms: 676K+ conversations, 25M+ tweets, 1,430 candidates, 5 months

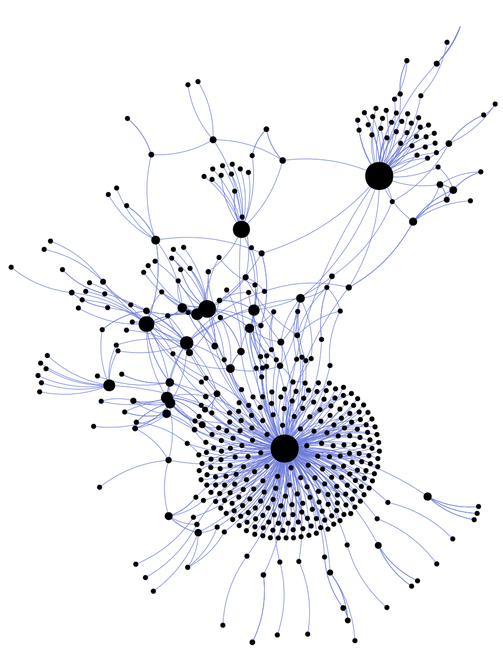
Representing A Twitter Conversation



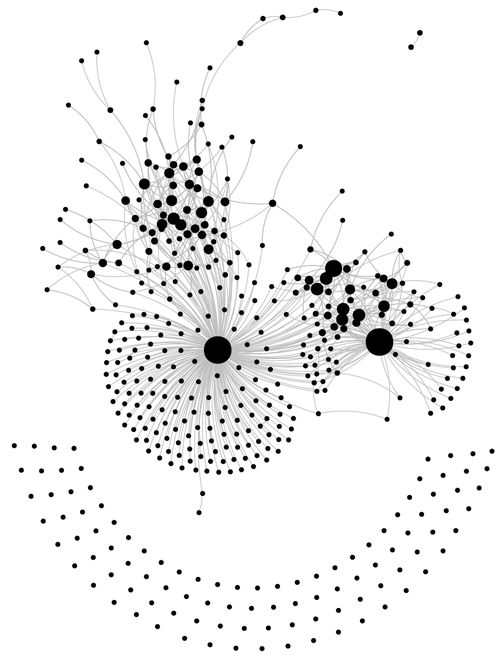
Twitter UI



Reply Tree



Reply Graph



Follow Graph

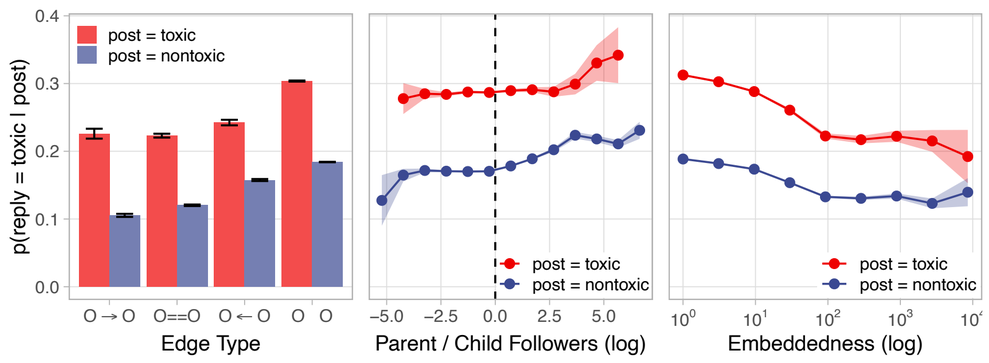
Analyses

Individual-level Analysis

- Toxicity is spread across many low to moderately toxic users

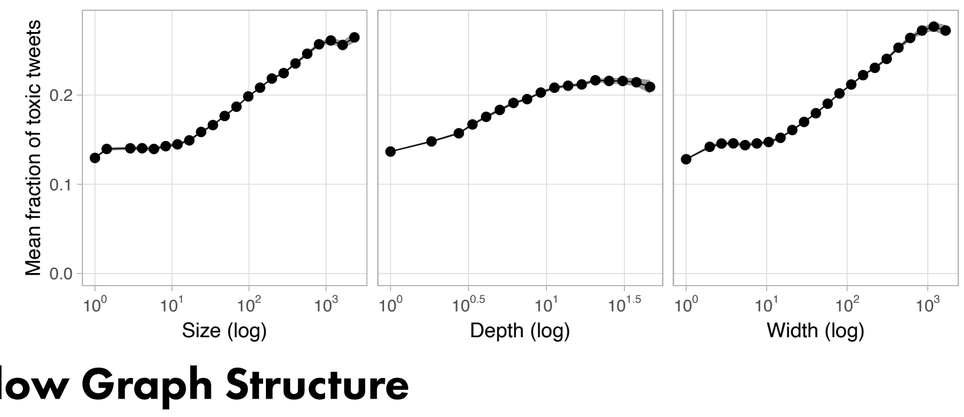
Dyad-level Analysis

- Toxic replies are more likely to come from other users who: (i) do not have any social relationship with the poster,(ii) have fewer followers, and (iii) do not have many common friends



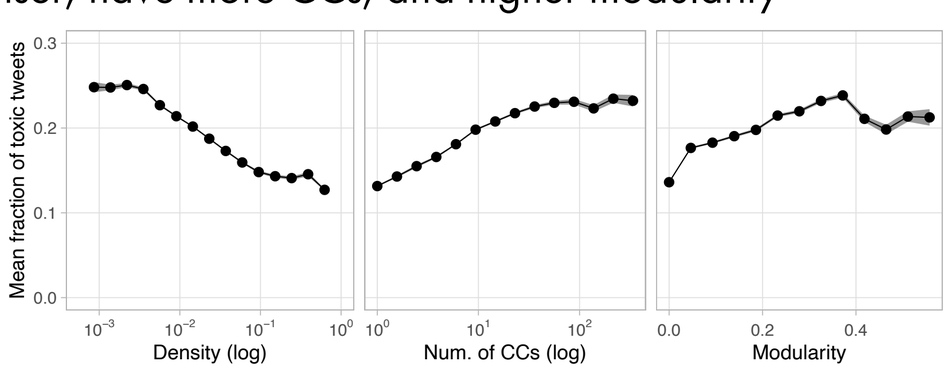
Reply Tree Structure

- Toxic conversations tend to have larger, deeper, and wider reply trees



Follow Graph Structure

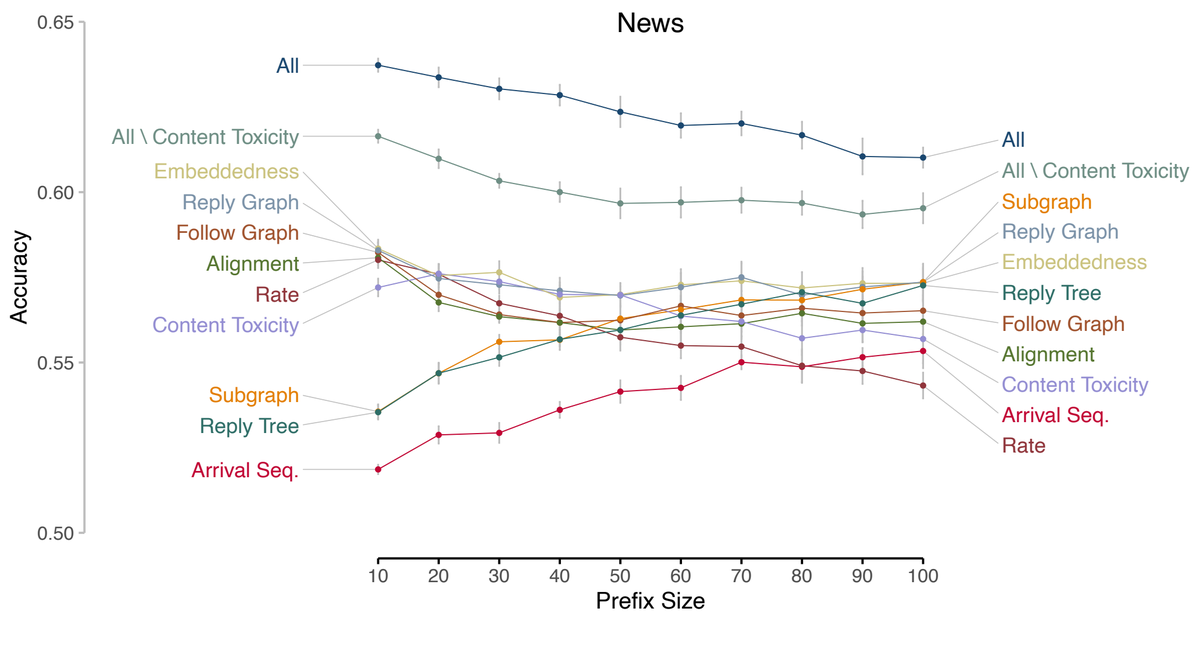
- Toxic conversations tend to have follow graphs that are denser, have more CCs, and higher modularity



Prediction

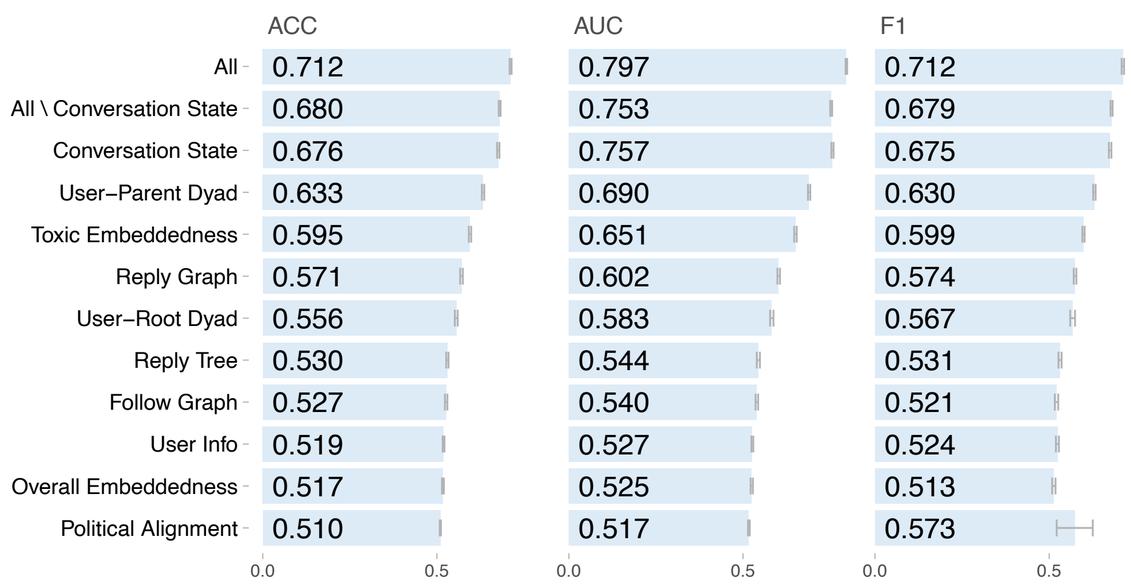
Future Toxicity Predictions

- Task: Given the conversation so far, predict whether the conversation will become more toxic than expected
- Using stratification to control for prefix toxicity



Next Reply Toxicity Predictions

- Task: User i is about to join the conversation, will they post a toxic reply?
- Paired prediction task to control for the root content



Extended version published at WWW'21: https://doi.org/10.1145/3442381.3449861