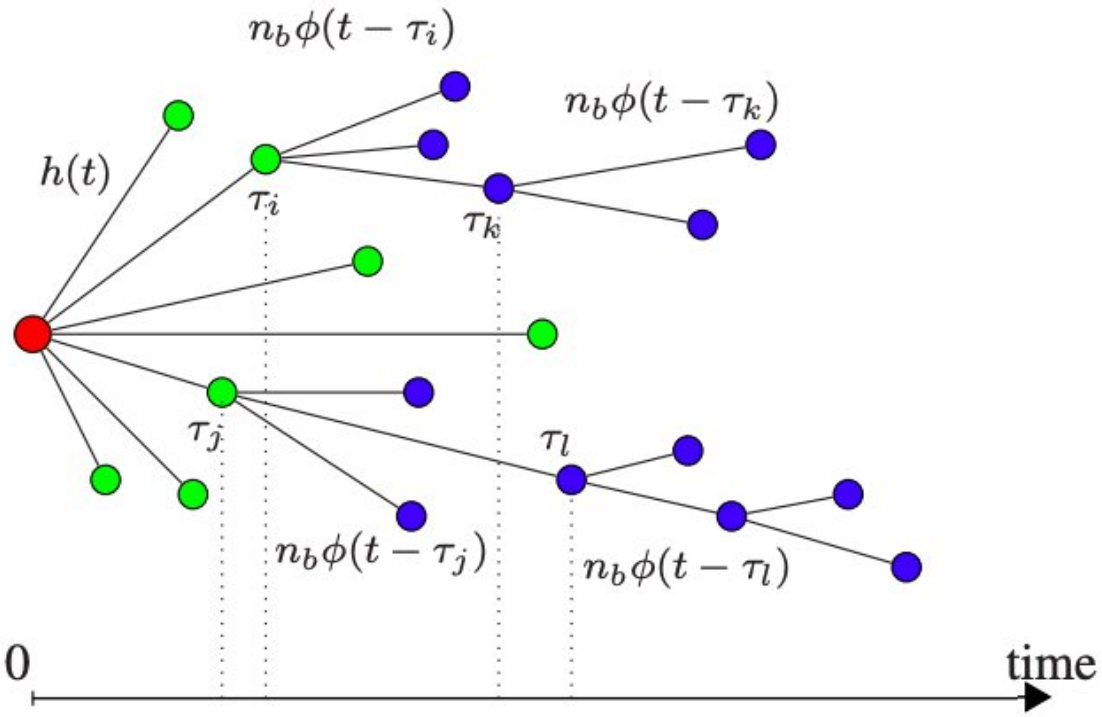
A model to predict the size and shape of online comment threads

November 5 2019, by Ingrid Fadelli



Example of a Hawkes branching process. The red node (far left) represents a social media post. Green and blue nodes represent 'immigrant' and 'offspring' events respectively. Credit: Krohn & Weninger, adapted with permission from the work of Medvedev et al.

On social media platforms such as Reddit and Twitter people can

express their opinions and take part in discussions about a variety of topics. This is generally done in comment threads, which allow users to comment on existing posts.

A comment thread is essentially a conversation between different online users in the form of comments. In computer science, comment threads are often regarded as "trees," with nodes representing the original post and subsequent comments, and directed edges representing "reply-to" relationships.

Two researchers at the University of Notre Dame have recently developed a model to predict the size and shape of online comment threads when viewing them as trees. They called this model, introduced in a paper pre-published on arXiv, the Comment Thread Prediction Model (CTPM).

"Our main research goal is to predict the size and shape of a comment thread on social media sites," Tim Weninger, one of the researchers who carried out the study, told TechXplore. "These sites allow users to post news or images or other content. Then other users like, share or comment on the post. We are interested mostly in comment threads, where a user can comment on the post itself or reply to comments like on Reddit and Twitter (but not Facebook or YouTube)."

The study carried out by Weninger and his colleague Rachel Krohn was funded by a US Defense Advanced Research Project Agency (DARPA) program, which specifically focuses on social simulation. One of the questions asked by this program is whether simulating social media activity is possible.

Previous studies suggest that the first few hours of a post's life are of vital importance in predicting its future popularity. In fact, posts that get a lot of early attention and are immediately commented on by users

generally spark further online discussion in the future. On the other hand, posts that initially do not receive much attention tend to also attract less attention in the future.

Most existing techniques designed to predict the size and shape of comment threads work by observing the first several comments that are added to a post and then creating a predictive model. However, as the majority of comment threads are relatively small, waiting for new data to be generated can impair the overall goal of the prediction task.

The DARPA program funding the study thus specifically instructed the researchers to investigate whether they could to predict a post's popularity, including the number of comments it would elicit in the future, based solely on its title. With this objective in mind, the team developed a model that analyzes the words in a Reddit post's title, along with the posting user and the subreddit to which it was submitted. These variables are used to create a "Hawkes process," a statistical model used to represent mathematical points in space.

"We use a Hawkes process to simulate how people view the post, read a comment, and then decide to reply to each comment," Weninger said. "The model isn't perfect and doesn't actually simulate the content of the comments (i.e. we don't guess what the comment actually says, just if there is a comment or not), however, on average we do a pretty good job at predicting which comments will be popular and which will not be popular just based on the title, author and subreddit of a post."

Weninger and his colleagues evaluated the CTPM model on thousands of real user discussions taken from Reddit, comparing its effectiveness in predicting the size and shape of comment threads with that of other techniques. Remarkably, their model significantly outperformed all the existing models and baselines that it was compared to.

"To me the most meaningful contribution of this work is the ability of our model to predict the size and shape of online conversations," Weninger said. "This is important to US law enforcement and defense agencies because being able to predict the future in cyberspace enables these agencies to prepare effective defenses against cyber-attacks and other events which frequently move from the cyber world to the physical world."

In the future, the model proposed by Weninger and his colleagues could be used to predict the popularity of posts on Twitter or Reddit based solely on their title. The team now plans to continue investigating how humans consume and curate information online, including their interactions with others' posts (e.g. likes, shares, retweets, etc.).

"The likes, shares, upvotes, and retweets provided by users are the single most important thing to social media companies because they indicate which content to promote and which content might be spam or low quality," Weninger said. "We study these processes and how they can be corrupted by individuals or groups with bad intentions. Our future work in this area will look at manipulations of social content (e.g. image alterations, photoshops, deepfakes, etc.), as we can learn a lot about people and their culture by watching how they alter images in social media."

More information: Modelling online comment threads from their start. arXiv:1910.08575 [cs.SI]. arxiv.org/abs/1910.08575

Modelling structure and predicting dynamics of discussion threads in online boards. DOI: 10.1093/comnet/cny010. academic.oup.com/comnet/articl … tract/7/1/67/4991998