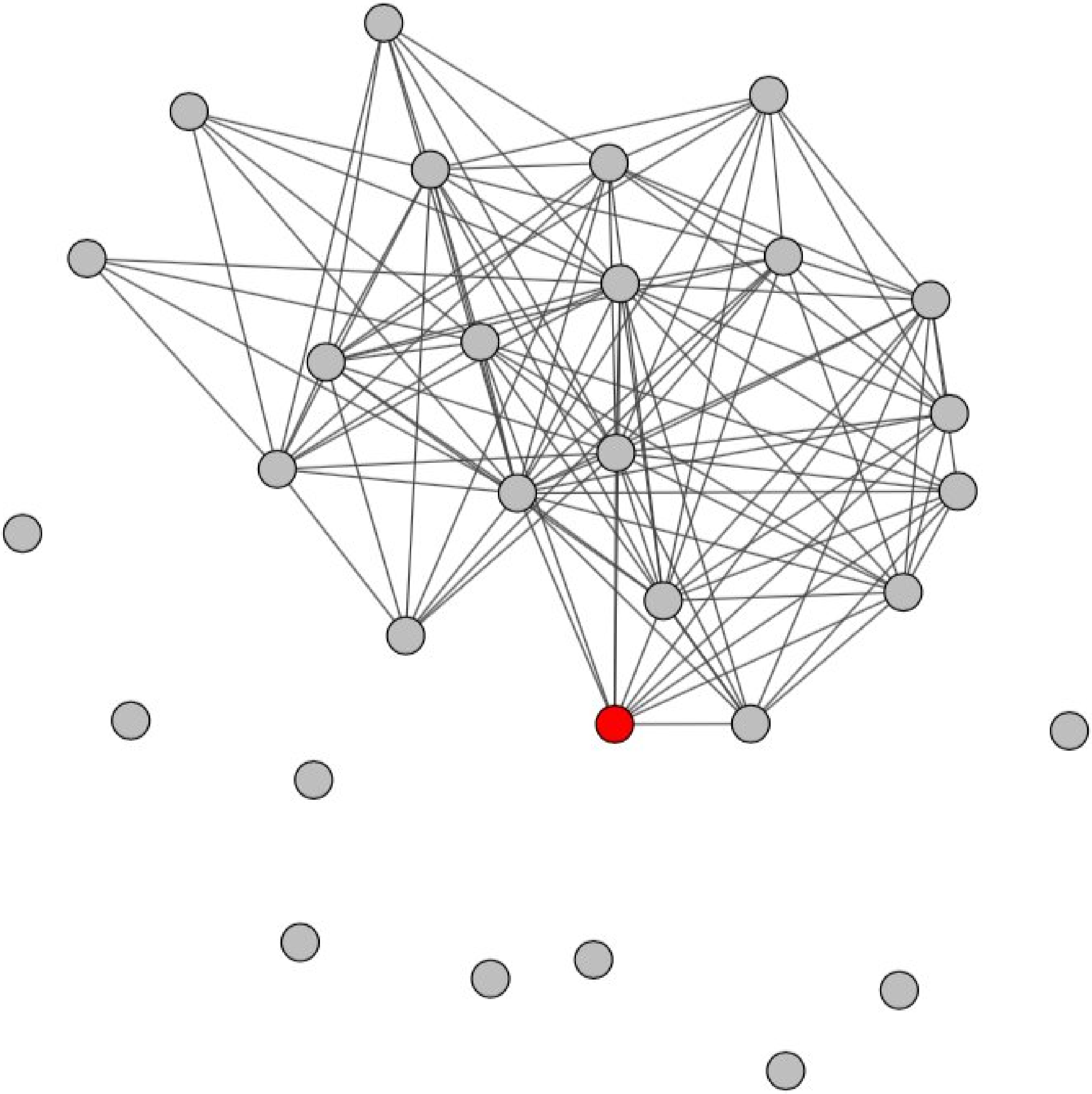
Researchers develop a new system to detect abuse in online communities

February 13 2019, by Ingrid Fadelli



Conversational graph obtained by considering a period of time preceding the abuse. Credit: Papegnies et al.

A team of researchers at Avignon University has recently developed a system to automatically detect abuse in online communities. This system, presented in a paper pre-published on arXiv, was found to outperform existing approaches for detecting abuse and moderating user-generated content.

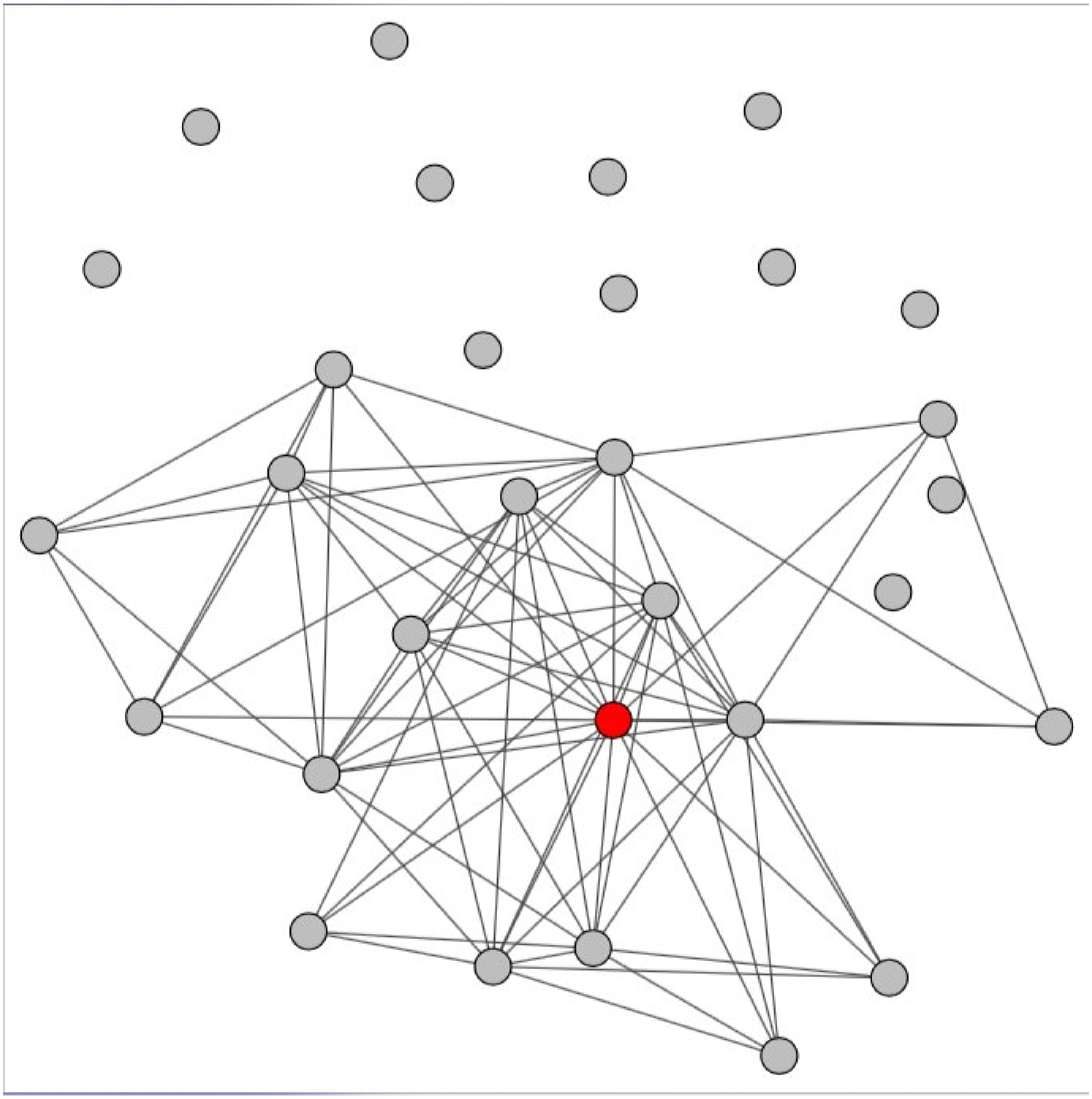
"Ever-growing online communities offer the opportunity to spread ideas through the internet, guaranteeing some anonymity to users," the researchers told TechXplore, via e-mail. "However, these spaces often have users displaying abusive behavior. For community leaders, it is important to moderate these malicious acts, as failure to do so could poison the community, trigger user exodus and expose administrators to legal issues."

The moderation of online user-generated content is generally carried out manually by humans; hence, it can be both expensive and time-consuming. To reduce costs, researchers have been trying to develop fully automated content moderation tools that could either replace or assist human moderators.

"In this work, we formulate the task of content moderation as a classification problem, and apply our method to a corpus of messages exchanged by players of an MMORPG, a massively multiplayer online role-playing game," the researchers said.

As a first step, the researchers extracted conversational networks from raw chat logs representing the conversations in which each abusive message was sent, and characterized them using topological measures.

They used their results as features, training a classifier to detect abuse on online platforms.

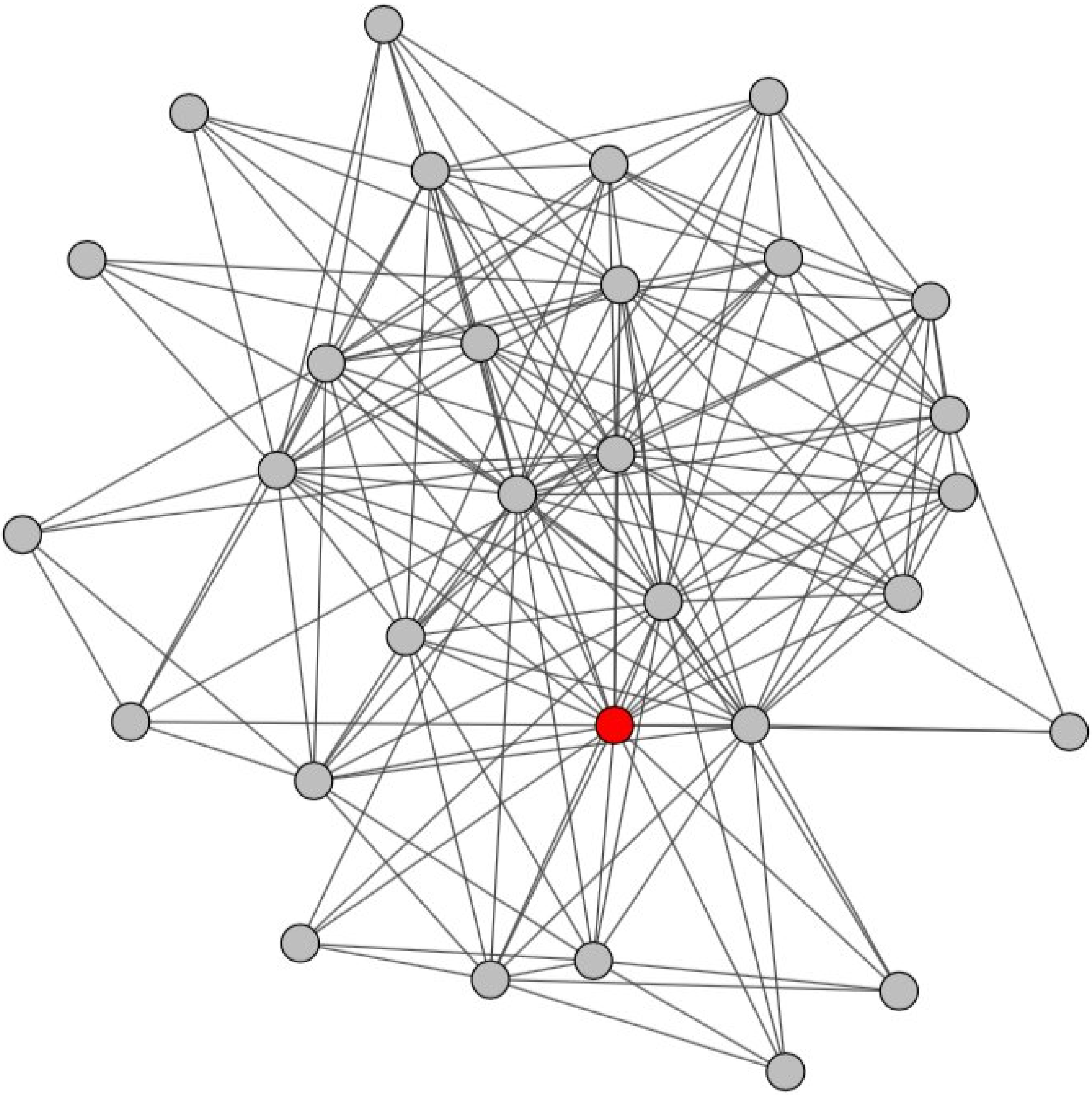


The conversational graph obtained by considering a period of time following the abuse. Credit: Papegnies et al.

When extracting the conversational networks, the researchers followed a three-step method. First, they identified the subset of messages that they would use to extract the network. Then, they selected a subset of users that were the likely receivers of each message. Finally, they added edges and revised their weights based on these potential message receivers.

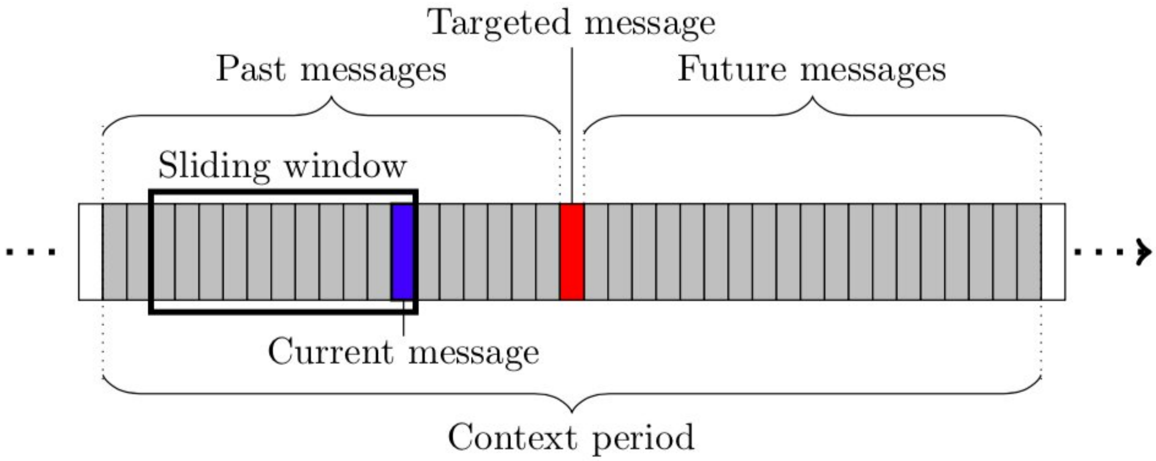
"Existing methods for the automatic detection of abusive messages focus on the textual content of the exchanged messages, which raises many issues: language-specific problems, syntax errors, spelling mistakes, obfuscation, and others," the researchers explained. "On the contrary, we use only the presence/absence of interactions between users, i.e. the fact that they exchange some messages (or not), by opposition to the nature of the exchanged messages. Ignoring the content allowed us to solve these issues."

Essentially, the researchers modeled online conversations using a graph in which nodes represent users and links represent message exchanges. Using graph-specific measures, they were able to observe differences in the way conversations are structured depending on whether or not they contain abusive messages. These differences were then used to train a classifier to detect abuse in conversations between users.

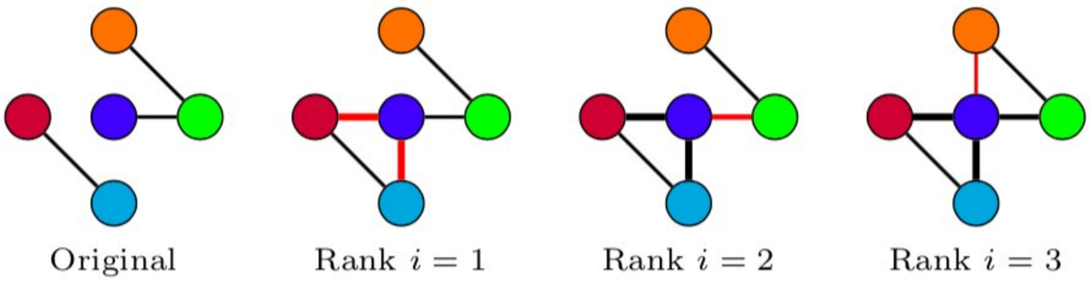


Conversational graph obtained by considering the whole period of time (i.e. both before and after the abuse). Credit: Papegnies et al.

"Our first effort, presented in a previous article, was based on the traditional approach, i.e., it used the textual content of messages," the researchers explained. "When we proposed this graph-based method, we did not expect it to work this well; we even thought that it would result in lower performances compared to the content-based method. We were very surprised to obtain significantly better results. This is the most meaningful finding of our study—that, at least for this specific task, the structure of the conversation is more discriminant than the nature of the exchanged content."



Credit: Papegnies et al.



Credit: Papegnies et al.

The researchers tested their system on a dataset of user comments from a French MMORPG game and found that it outperformed existing approaches, with an F-measure of 83.89 when using the full feature set. By reducing the feature set and keeping only the most discriminative features, they were able to dramatically reduce computing time, while retaining excellent performance. In the future, their graph-based approach could also be applied to other message classification tasks, such as online troll detection.

"We will now try to merge both approaches (content- and graph-based), in order to check whether they take advantage of similar information, in which case the results would be similar, or if they rely on complementary information, in which case, combining them should lead to improvements in performance," the researchers added. "Then, we want to move towards a more automated method to characterize our conversational graphs, called graph embeddings. It is a deep learning based method that consists in training a neural network to get an efficient representation of the graphs. By comparison, we currently do this part of the work manually, via a task called feature selection."

More information: Conversational networks for automatic online moderation. arXiv:1901.11281 [cs.IR]. arxiv.org/abs/1901.11281

Impact of content features for automatic online abuse detection. DOI: 10.1007/978-3-319-77116-8_30. link.springer.com/chapter/10.1 … 978-3-319-77116-8_30

Citation: Researchers develop a new system to detect abuse in online communities (2019, February 13) retrieved 9 April 2024 from https://techxplore.com/news/2019-02-abuse-online.html