# Predicting structure and dynamics of discussion threads in online boards using Hawkes processes

Alexey Medvedev[1,2] and Renaud Lambiotte[1,3]

[1] NaXys, Université de Namur, Namur, B-5000, Belgium
[2] ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, B-1348, Belgium
[3] Mathematical Institute, University of Oxford, Oxford, UK
an_medvedev@yahoo.com,
WWW home page: http://alexeymedvedev.com

Online social platforms provide a fruitful source of information about social interaction. Depending on the platform, various tree-like cascading patterns emerge as a consequence of such interaction. For example, on Twitter or on Facebook people interact via resharing messages, which turns into cascade trees of reshares, in email networks people forward messages to their peers resulting in trees of email forwards, in online boards like Digg or Reddit people interact via discussing particular posts, which leaves a trace of discussion trees. The two main questions arise: what is the shape of these cascades and how to predict the dynamics of their evolution?

The question of evolution of discussion threads is now gradually being understood. In [1, 2] the authors studied only the structural evolution of discussion trees in four large Internet boards, and they suggested a tree generation model based on preferential attachment (PA) mechanism. However the dynamical properties are left out of consideration. In [3] the authors introduce a merely theoretical model which aims to describe structural and temporal evolution of the discussions. Their proposition is to use a specific Levy point process to generate timings, then construct the PA discussion tree assigning to each new node a subsequently generated timing. However, being a sort of a mean-field model, it describes evolution on average, thus having limited utility in practice.

We consider cascades given by discussion trees of posts in online board Reddit. The dataset of Reddit discussion threads consists of all posts and comments submitted to Reddit from Jan, 2008 till Jan, 2015. The dataset in total contains more than 150 million posts and around 1.4 billion comments. We propose a model of discussion trees generation based on the self-exciting Hawkes processes, which represents both the tree structure and temporal information. We use the dataset of Reddit discussion threads to show that structurally trees resemble Galton-Watson trees with a root bias, and distinct the cases when the dynamics of comments attraction can be well predicted using Hawkes processes.

## References

1. Gómez, V., Kappen, H.J., Kaltenbrunner, A.: Modeling the structure and evolution of discussion cascades. In: Proceedings of the HT '11. pp. 181–190 (2011)
2. Gómez, V., Kappen, H.J., Litvak, N., Kaltenbrunner, A.: A likelihood-based framework for the analysis of discussion threads. World Wide Web 16(5-6), 645–675 (2013)
3. Wang, C., Ye, M., Huberman, B.A.: From user comments to on-line conversations. In: Proceedings of the KDD '12. pp. 244–252. KDD '12 (2012)