PROJECT SUMMARY - GROUP 32

Story

Until recently, researchers have mostly relied on classical publication tools to share their work within their field. With the increased popularity of social media, however, researchers have been starting to use Twitter to further enhance their impact across the scientific community. However, it is not clear whether this increase in Twitter usage is just following a general trend in society or if it represents a specific strategy that turned out to be useful for publication success. To tackle this question, our project aims, in a first exploratory phase, at understanding the structure and behaviour of this very specific, newly emerging network of scientists on Twitter. In a second phase, we seek to relate this Twitter activity to activity outside of the Twitter network, for example to predict publication success.

Data Acquisition

We intend to use a dataset from Hadgu & Jäschke (2014), which puts together over 9000 computer scientists' Twitter accounts from an AI conference list and links them to features about Twitter behaviour as well as to an academic profile page. We will extend this database by crawling several time-series of Twitter activities by using the Twitter API. For the second part of the project, based on the academic profile pages, we will extract data from Google Scholar, such as the h-index, a list of co-authors, as well as time-series of publications and citations, using a python library for web-crawling.

Exploration

As a first exploratory phase, we will simply build an epsilon-similarity graph based on the features of the Hadgu & Jäschke (2014) dataset and analyse its basic properties, such as degree distribution, diameter or clustering, and compare those properties to different existing network models. Next, we will build a second epsilon-similarity graph based on the similarity of the previously crawled time-series of individual Twitter activities, to see if we can identify, for instance, hypes for certain subfields within computer science.

Exploitation

To probe the importance of Twitter activity in science, we seek to test the predictive power of Twitter-behaviour-related features with respect to information that we crawl from outside the social network, such as the h-index and time-series of publications or citations. To this end, we intend to extract relevant features from the data by using graph filters, whose parameters would be trained using graph neural networks, based on prediction performance.

To help deal with the expected confound that successful computer scientists may simply be more popular on Twitter, even if only sporadically using the platform themselves, we intend to build a graph from the Granger causality matrix between Twitter and Google Scholar time-series (Fig. 1). This would help us to determine whether Twitter time-series influences Google Scholar time-series or the other way around.

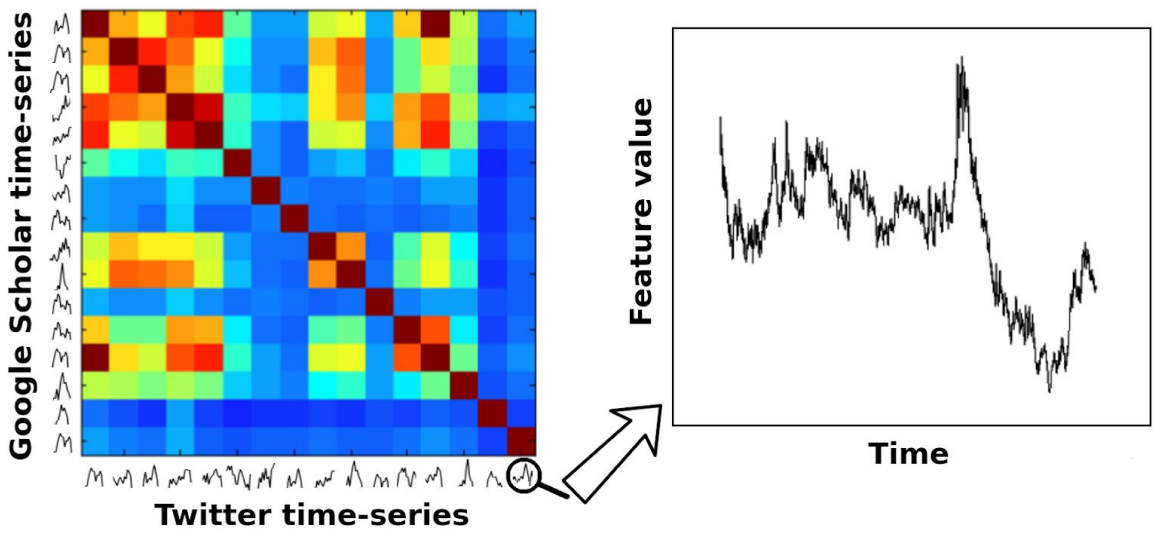

Figure 1. Left. Illustration of a Granger-causality matrix that we could build, comparing different Twitter and Google Scholar time-series from computer scientists that use Twitter. This matrix can be thresholded and used as a graph for further knowledge. **Right**. Zoom-in on how any time series might look. The monitored feature could be, for example, the amount of followers a user gets every month on Twitter.