École Polytechnique Fédérale de Lausanne

Network Tour of Data Science

EE-558

Term Project Final Report

Group 32

STUDENT ID :
Alban Bornet
Gizay Ceylan
Wei-Hsiang Lin
Lukas Vogelsang

Contents

1 INTRODUCTION

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 whether it represents a specific strategy that turned out to be useful for publication success. In the first part of this project, we conducted a basic assessment of this newly emerging network of scientists on Twitter – by means of similarity graph analyses and spectral clustering-based predictions of features concerning Twitter behavior reported for several thousands of computer scientists, in Hadgu and Jäschke (2014). In the second part, we tried to relate this Twitter activity to activity from outside of the Twitter network. Specifically, we probed whether time-series of the activity of the computer scientists' Twitter activity is a good predictor of either their $h - index$ or of the number of citations they get every year on Google Scholar. To gain predictive power, we filtered the data using a network built from co-author relationships between the scientists, as the underlying structure of those time-series.

2 PART I: Basic analysis based on a data from Hadgu & Jäschke (2014)

In part I, we utilized a dataset from Hadgu Jäschke (2014), which puts together several thousands of 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.

2.1 Data Acquisition

We selected the set of Twitter accounts that were listed in all relevant sub-tables and merged the data into a single database, holding a set of features for a total of 8605 Twitter accounts of researchers.

2.2 Exploration

As the first step of our exploratory analysis, we constructed a similarity graph based on the (thresholded) Euclidean distances between 17 (normalized) Twitter-related features we deemed meaningful (the full list can be found in the documentation of the GitHub repository). The resulting graph, holding 8605 nodes, 9281909 edges 0 self-loops, and 55 connected components, interestingly, shows a bimodal degree distribution (see Figure 1A). This means that some researchers are very similar to others in their Twitter behavior, while others are very different. However, it is important to acknowledge here that the number of features utilized for the similarity graph is limited and that several Twitter users are fairly inactive. To probe the general usefulness of utilizing our set of 17 features, selected and recombined from the Hadgu and Jäschke (2014) database, we first calculated the features' correlation matrix. While features such as the number of hashtags and number of conference hashtags were highly correlated, many correlations were close to zero (see Figure 1B). For example, contrary to the intuition that researchers with many publications would have many followers, the correlation coefficient was -0.0064. There are many possible confounding factors that limit the interpretability of this value in the scope of assessing the usefulness of Twitter to predict publication success: for instance, researchers with many publications could, on average, be older and, on average, less drawn to using Twitter as a scientific platform.

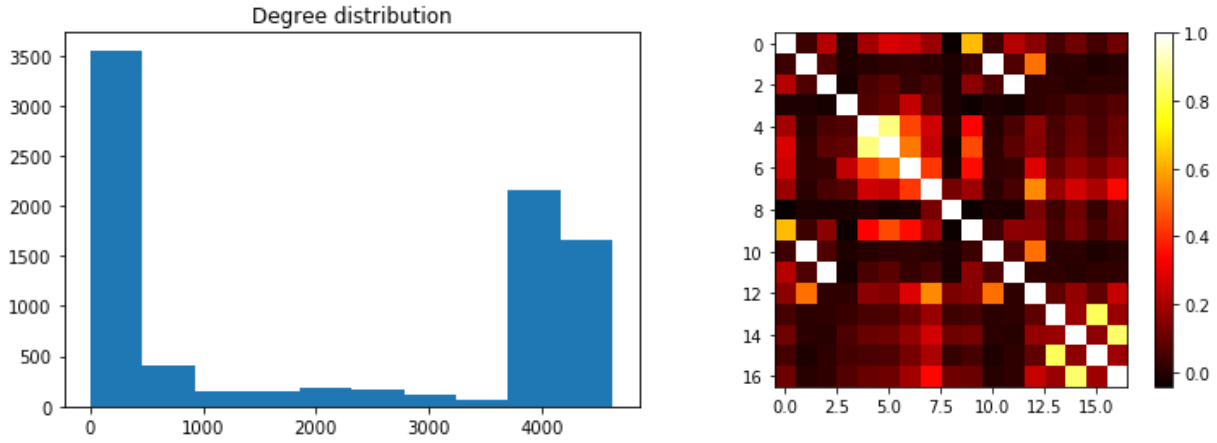

Figure 1: Left: Degree distribution of feature similarity graph visualized Right: Correlation matrix between static Twitter features visualized

2.3 Exploitation

While the correlation matrix of single Twitter features revealed limitations to our investigation, we were able to show some degree of predictiveness. Specifically, we constructed an epsilon-similarity graph based on the Twitter features (with sigma calculated as the standard deviation of the network edges weights; epsilon set to zero to ensure a single giant component) and employed a spectral clustering approach, based on the spectral decomposition of the resulting adjacency matrix, to predict the labels 'male' vs 'female' as well as the labels 'phd' vs. 'prof'. Note that we carried out two separate analyses and only included the researchers with documented ground truths (i.e. documented gender or academic position). Using this analysis, we were able to show that the academic position could be predicted more easily than the gender from the Twitter data: our gender prediction performed rather modestly with a precision, recall, and f1-score of 0.83, 0.72 and 0.77 for male researchers ($n = 4147$) and 0.22, 0.35 and 0.27 for female researchers ($n = 953$) but our position prediction performed above chance level with regard to all relevant metrics (precision, recall, and f1-scores of 0.75, 0.68, 0.72 for PhD students ($n = 911$) and 0.56, 0.64, and 0.6 for professors ($n = 570$)). Note that we were able to replicate the success in prediction when excluding the 'publication' non-Twitter feature of our database (all other features were Twitter-based). It ruled out the possibility that the results were biased since the publication itself had a strong relationship with the academic position. While this tendency lends support to the proposal that academic success can be predicted from the Twitter features to some degree, utilizing this database of static Twitter features did not allow to account for many of the confounding factors. We hoped that utilizing temporal features in combination with predictive algorithms and causal analysis tools could help alleviate some of those concerns.

3 PART II: Twitter and Google Scholar time-series analysis

In this part of the project, we extended the original dataset, starting from the twitter accounts of the scientists and their academic (dblp) profile page, to retrieve Twitter activity time-series as well as time-series based on the number of citations per year and $h - indexes$ from Google Scholar. We then used those data to test the predictive power of Twitter activity on publication success.

3.1 Data Acquisition

First, using the Twitter API, we extracted time-series of the number of Twitter posts written every month by each scientist. We retrieved all the tweets (including tweets and retweets) from all the

scientists. We then created the time-series based on the month of each post. Furthermore, using a Python library for web-scraping and the list of academic profile page URLs associated with the scientists in the original dataset, we automated the extraction of publication-related information from their personal Google Scholar profile page (if existing). This process was particularly tedious because we had to make many requests to Google Scholar in a short amount of time. To avoid overloading the servers as well as to avoid Google Scholar sending the "429 - Too many requests" error, we set up random delays between each request. Moreover, every time a captcha page arose instead of the Google Scholar page, our algorithm automatically changed the IP address, using NordVPN. After one week of web-scraping, the algorithm collected the number of citations that all scientists in the original dataset got each year, their $h-index$, and a list of their co-authors.

3.2 Exploration

To probe the causal influence from Twitter activity to publication success, we measured the Granger causality between each pair of Twitter and Google Scholar time-series available in our scraped database (keeping only the pairs with a $p-value$ of 0.01 or less). The resulting matrix and the associated degree distribution is shown in Figure 2.

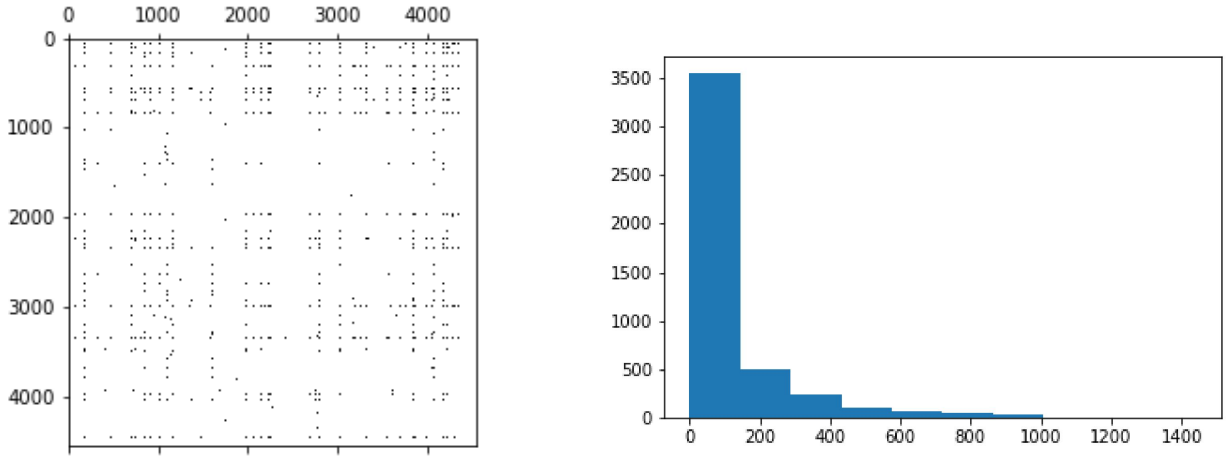

Figure 2: Left: Granger causality matrix from Twitter activity time-series to Google Scholar citation time-series Right: Degree distribution of the Twitter activity matrix

A visual inspection of the matrix (see Figure 2A) revealed that Twitter time-series of some researchers might have an effect, not only on the Google Scholar time-series of themselves but also of other researchers' time-series. To probe whether those researchers happened to be the co-authors or otherwise influential in their Twitter behavior, we chose researchers within the top 50 degrees and tested whether they are the head of the academic. We found that the mean $h-index$ of the researchers within the top-50 degrees is 25.86, whereas the mean $h-index$ for the entire dataset is only 8.78. This places the top-50 degrees researcher in the top 10% of all researchers measured by their $h-index$ (the 90th percentile of the researchers' $h-index$ is 25). This indicated that the graph helps to identify influential researchers in the field, and also motivated the exploitation phase.

3.3 Exploitation

Since Twitter time-series seem to have an effect on publication success, we tried to directly probe the predictive power of Twitter monthly time-series. To this end, we fed the Twitter data to a classifier trained to predict either the scientist's $h-index$ or the Google Scholar time-series.

To help the classifier, we hypothesized that the underlying structure where all those features evolve is the "space" of scientists being related to each other through co-authorship. For example, when a paper is out, all co-authors would co-jointly promote the publication on Twitter. For this reason, we built a network by linking any pair of scientists that are both in the original dataset and that are co-authors. Then, we used it to build a graph filter that would help the classifier to select what part of the input Twitter time-series is more useful to predict publication success (see Figure 3). Unfortunately, the co-authorship graph was very sparse and hence was very far from being connected (many eigenvalues of the Laplacian are zero). We originally wanted to "augment" the graph by adding connections between scientists that follow each other on Twitter, but the Twitter API was too restrictive to allow us to get those data on time.

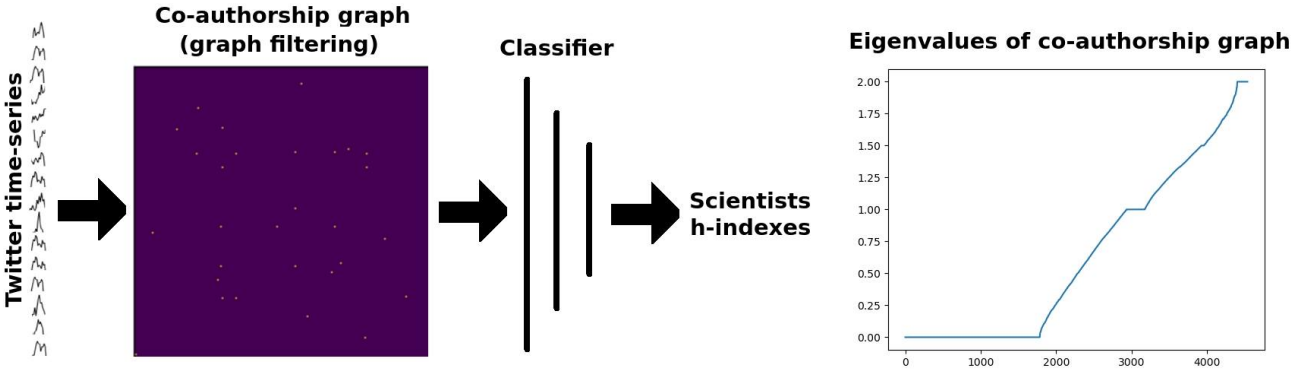

Figure 3: Left: Illustration of the method used to predict publication success from the Twitter time-series Right: Eigenvalues of the co-authorship graph

To build labels from the $h-indexes$, we binned those numbers in 15 different classes (class n would correspond to all the scientists that have an $h-index$ between $n*10$ and $(n+1)*10$). The classifier is a set of 1D convolutions that ends up with a fully connected layer. After training the classifier for 200 epochs, either using or not using the co-authorship graph as a feature graph filter, we obtained testing accuracies of, respectively, 33% and 38% (chance level is 1/15, which is approximately 6.7%). The evolution of the loss and hit rate for the training set and the validation set was also different between both conditions (see Figure 4).

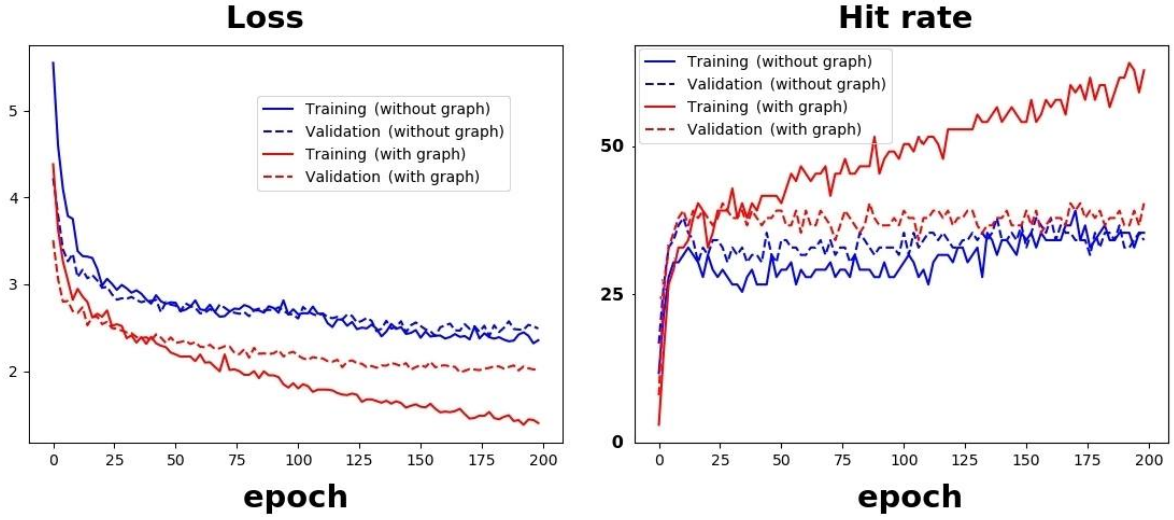

Figure 4: Evolution of the loss and hit rate of the $h-index$ classifier during training

To try and predict the number of citations per year from Twitter activity, we planned to use a network of LSTM units, which receives the Twitter data and then sends the output to a set of 1D convolutions that reduce the dimension of the output to the number of years contained in the year Google Scholar

time-series. However, this turned out to be over-ambitious, given our knowledge of deep learning, the time we had, the size of our dataset and the sparsity of the coauthorship matrix. Hence, we don't report any results of this analysis.

4 DISCUSSION

In the context of increased usage of Twitter as an academic platform, we here set out to probe the predictive power of both static and temporally-evolving Twitter features on academic attributes.

In the first part of the project, we found that several Twitter features were not as predictive of other features as we would have expected. For instance, we observed the absence of a correlation between the number of citations and the number of followers; possibly biased by confounding factors. However, combining several features and applying a spectral clustering approach, we were able to predict, better than chance level, whether a set of researchers would be a PhD student or a professor.

In the second part of the project, we specifically investigated the potential effect of Twitter activity on publication success. In the first exploration phase, we used Granger causality and found that the Twitter activity of some researchers has a strong influence on other researchers' Google Scholar time-series. By extracting information within the Granger causality graph, we found that the researchers with high degrees were also those who had high $h-indexes$. This result hinted to us that Twitter activity might have some predictive power for publication success. In a second phase, we tested this predictive power by training a classifier to predict scientists' $h-index$ from their Twitter activity time-series. After training, the classifier performance was higher than the chance level. Interestingly, using graph filters, adding network knowledge about scientists' co-authors helped the classifier to use more relevant Twitter-related information and increase its predictive power. This was to expect since those Twitter signals are supposed to evolve in a space that is shaped by co-author relationships. It would have been even more interesting to use network information about which scientists follow each other on Twitter because Twitter time-series truly evolve in this space. Lastly, we were unable to reconstruct Google Scholar time-series from Twitter activity, but we believe that a more sophisticated and extended analysis would lead to some interesting results.

Public Github repository: `https://github.com/albornet/ntds_2019_team_32/`

5 REFERENCES

[1] Hadgu, A. T., Jäschke, R. (2014, June). Identifying and analyzing researchers on twitter. In Proceedings of the 2014 ACM conference on Web science (pp. 23-32). ACM.