Creating hyper-realistic digital twins of social networks by combining data mining, natural language processing and agent-based modelling

Max Rollwage[1*], Perukrishnen Vytelingum[1]
[1]Simudyne, UK; *corresponding author: max@simudyne.com

Keywords: social networks, spread of (mis-)information, NLP, behavioral modelling, fake news

Background

Social media has become a primary source of information and is the “go to” place for people’s daily news consumption. Understanding how information spreads on these platforms is a critical research endeavour with applications ranging from online marketing to the spread of misinformation or the formation and burst of social network driven financial bubbles (e.g., GameStop). Compared to traditional news outlets, the distribution of information on social media is decentralized and depends on the sharing behaviour of every single entity in the network. Thus, the spread of information online is an emergent property of agents’ micro-actions. Therefore, agent-based modelling is an ideal approach for modelling the dynamics of information spread in social media networks.

However, while ABMs have a clear appeal in this situation (as they allow to exactly model the underlying processes), their usage has often remained somewhat theoretical/abstract. A potential reason for this is that ABMs often model abstract populations of agents that resemble real-world population with respect to some summary statistics (e.g., average node degree and assortativity) but do not directly map onto any real-world entities. Thus, agent-based modelling has been useful for understanding general rules about how information spreads in social networks but has (to our knowledge) rarely been used to make predictions about specific real-world situations.

Here we present a workflow to address this, which entails data mining from social media platforms, natural language processing, behavioral modelling, and agent-based simulations. This workflow enables to create hyper-realistic digital twins of real-world social networks. Importantly, we show that this level of realism leads to highly accurate predictions about how and which information will spread in a social network of interest, outperforming other machine-learning techniques.

Methods

As example data set, we recreated the first author’s follower network on Twitter[i] and simulated how their tweets are shared/retweeted in the network. For this purpose, we made heavy use of Twitter’s developer API for scraping/mining data from this social media platform.

First, we acquired data about all the author’s followers and their interconnections, enabling us to recreate an exact copy of the social network. This network focused on direct followers but did not model the potential sharing of tweets of more distant agents (e.g., followers of followers).

Second, for each agent we inferred relevant attributes that determine whether they are likely to share/retweet a specific post. Hereby, we focused on the agents’ interests and their general activity level on Twitter. This was done by scraping each agents’ timeline (i.e., list of previously shared posts) to infer the frequency with which they generally share posts. Moreover, we applied natural language processing to infer each agent’s interest in specific topics, specifically focusing on the overlap between each agent’s interests and the posts tweeted by the author. This was achieved by transforming each post into a vector using a TF-IDF representation[ii] and then calculating a cosine similarity between tweets:

$$similarity = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are components of the vector A and B, whereby each vector represents a tweet.

Finally, we derived a list of people who have shared/retweeted each of the authors posts. Having inferred each agents’ attributes (interests and activity level) and having data on whether they have shared certain tweets, it is

[i] The author analysed their own private network to avoid any privacy issues that could potentially be associated with scraping and analyzing someone else’s data (although Twitter’s privacy policy would allow that).
[ii] Term Frequency - Inverse Document Frequency representation in scikit-learn

possible to fit a logistic regression model to predict sharing behavior based on people's attributes, i.e., calibrating a link between attributes and sharing behaviour.

Bringing all these components together, we implemented a highly realistic virtual copy of the social network in an ABM whereby each agent represented a real-world twitter account with an exact recreation of connections between accounts, empirically derived attributes, and calibrated behavioral functions. Using this model, we aimed to simulate the spread of an average tweet posted by the author, using the number of retweets as output measure from each simulation run. We simulated 1500 Monte Carlo runs of the ABM simulation to derive a distribution over the expected number of retweets which can be compared against the empirical distribution of retweets observed on the author's account. Calculating the Kullback-Leibler divergence between these distributions can be used to formally evaluate the goodness of fit of the model.

For benchmarking the ABM further, we compare its goodness of fit to a set of alternative machine-learning approaches for predicting the number of retweets. Hereby, the alternative approach with best predictive power also modelled choice behaviours at the level of each agent. This model was based on the logistic regression model outlined earlier but instead of using a binary cut-off for deriving predictions, the class probability (probability to retweet) was used as input to a Bernoulli distribution and a sample drawn from this distribution determined the model prediction. This procedure ensured that model predictions accurately represented each agent's actual probability to retweet a specific post.

Results

The results indicate that the ABM model captures the empirical data very closely (see Figure 1A), suggesting that the very high level of realism in this model indeed enables a highly accurate model fit. Importantly, the ABM approach clearly outperforms even the best alternative approach (i.e., probabilistic logistic regression), providing a significant better model fit (Wilcoxon signed-rank test, $p = 0.002$, see Figure 1B). The probabilistic logistic regression mainly fails to fit the empirical data by missing its long tail. This is likely been driven by the lacking network structure, not modelling network effects of repeated circulation of a post in the network. To test this intuition we systematically reduced the ABM's connectivity, showing the ABM converges to the results of the probabilistic logistic regression when interconnections between followers are omitted. This result further shows the critical importance of explicitly modelling network structure when simulating the spread of information on social media.

Discussion

We have outlined a workflow to implement empirically derived social networks in ABMs and through this create highly accurate simulations for the spread of information in these networks. Importantly, a similar approach can easily be implemented for topics of more theoretical interest such as the spread of misinformation or the formation and burst of financial bubbles driven by social media and internet forums. The focus of this paper is the general approach of how to derive empirically informed ABMs from social media data. This toolkit enables the creation of highly valuable models for a wide range of topics and applications related to information spread on social media.

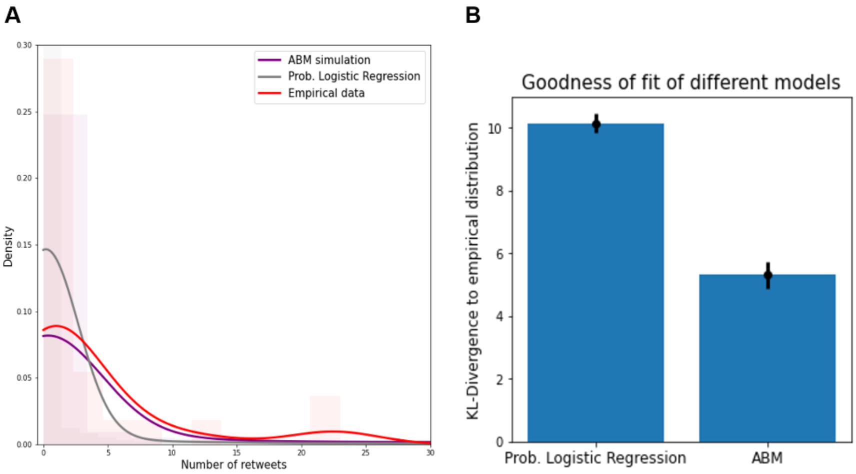

Figure 1. Comparison of empirical data and model predictions from the ABM and a probabilistic logistic regression. **A** Distribution over expected number of retweets. The distributions for the ABM and the prob. logistic regression are derived from 1500 Monte Carlo runs. **B** Kullback-Leibler divergence between empirical distribution and simulation results. Lower values indicate better model fit. Error bars represent standard deviation of 10 repititions of the 1500 Monte Carlo runs for deriving simulated distributions for each model in order to assess that the models produce consistent and repeatble predicitions.