Complex Networks, 2022

Simulating news spreading with an epidemic model on news network

Daniele Buschi, Lorenzo Sani

February 28, 2022

Abstract

In this project we implemented an epidemic-like model to study the news spreading on a network of articles relative to a specific story. We extended the analysis of the *hyper parameters* to a model proposed in 2018 in order to characterize the stories that have been diffused. In this report we will describe the data used for this analysis and methods we used. We will propose a more deep discussion about the *hyper parameters* of the model we implemented.

Contents

1 Scope

The purpose of this project is to study the global spread of the news using many channels available nowadays, e.g. newspapers, magazines, internet sites, social media, and many others. During last few years, due also by the infodemic caused by information related to COVID-19 pandemic, the global news spread has shown its properties in the velocity of the spreading and the global connections of sources. Then our interests in studying the dynamic of news spreading motivate this work. In this project we aim to model the spread of any kind of news to show some general behaviours on how specific articles' publication can influence, under an *"epidemic-like"* diffusion, other articles from all over the world. The basic idea underneath this work is that whatever is the story to tell, it is always possible to find some *macroparameters* which are able to describe the dynamic of its diffusion.

1.1 Dynamic of the system as epidemic process

We implemented a model firstly proposed in [1]. There, the authors studied a procedure to build a news spread network about a specific story and then study the diffusion of the story using an epidemic Susceptible-Infect-Recovered (SIR) model applied to that network. Even though several possible causes could determine the different spread of a specific story, e.g. the reputation of the original publisher, the importance of the story or the time of the publication, the dynamic of its spread have general properties. These properties can be studied applying concepts of temporal networks and topic modeling, connecting similar articles within the bounds of temporal window of influence. Therefore the choice to use an epidemic SIR model to describe the network's temporal evolution does not seem unreasonable and, as proven in [2], studying the dynamical processes of epidemic spreading in complex networks is central and it gives great results for time-varying networks. In this work we aim to study the application of this model for different stories using articles from all over the world. We additionally have studied a different process for initializing the *hyper parameters* of the model. The final objective is to propose a categorization of stories on the basis of the values of these *hyper parameters.*

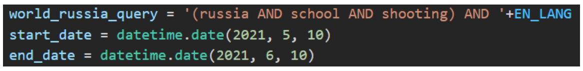
2 Data

We used data obtained by Media Cloud project [3]. Media Cloud is an open source and open data platform for storing, retrieving, visualizing, and analyzing online news. One can find out how much the media have been talking about your subject of interest over time, which were the key events that drove coverage about it, which are the words most frequently used around the keywords you searched for, and which media sources have covered the issue. Exploiting this features is possible to extract even data about the words of the articles, but not the entire article text because of copyright concerns. Data extraction is based on database queries, we exploited the *Python* APIs provided by Media Cloud project *GitHub* repository. Tutorials on the usage of these APIs are provided at [4]. Media Cloud provides also internet browser tools, such as Topic Explorer and Topic Mapper. Querying properly the database is fundamental for work, since the model we implemented is suitable to specific types of stories. In fact the story must satisfy the following properties:

- there must exist a sufficiently high number of articles which talk about the story;
- the temporal spreading of the story must be closed, meaning that after some time the story has no more audience and no more new articles that talk about that are published;
- there must exists a few articles published at the beginning, treated as initial infected.

Usually stories that satisfy these conditions above are serious crime news or they are about an event that has had a global resonance.

For our purposes 3 very famous event has been chosen: the Capitol Hill mob, the school shooting in Kazan (Russia), and the summer camp massacre in Norway. For each of these events, we built queries for Media Cloud APIs. An example of these is shown by Fig. 1, where starting and ending dates of the articles' search have been set along with some key words for the story.



```
world_russia_query = '(russia AND school AND shooting) AND '+EN_LANG
start_date = datetime.date(2021, 5, 10)
end_date = datetime.date(2021, 6, 10)
```

Figure 1: Capitol Hill query for Media Cloud API.

3 Methods

The model proposed in [1] is based on the construction of the network of articles about a story. For this specific construction, they chose a method to measure how similar articles are between each other. The fundamental tool for doing this comparison is the so-called *Word Similarity Matrix* from natural language analysis. Another important parameter for building the reconstructed network is the influence time window, i.e. the temporal distance threshold between two articles for building a link. The epidemic counterpart of the time window is the time of infection. The diffusion model is then based on a classical SIR epidemiological dynamical model upon the reconstructed network, whose only hyperparameter is the transmission parameter.

3.1 Similarity between articles

The measure of similarity between 2 articles has extensively been studied ([5]). According to the literature, in order to capture the semantics of the articles, one builds a word vector representation for every word in the corpus' vocabulary, taking into account the co-occurrence of words within a sentence.

For the generation of the *Word Similarity Matrix* we get, from Media Cloud, the information about the words and their frequencies in the articles queried. From these we extracted a data frame in which a raw represented the articles queried, a column represented a word, and a cell contained the counting for the word in the article. With this extraction we are then focusing only in the word used neglecting higher level linguistic features of texts such as syntactics or semantics.

In order to capture the semantics of the articles, the data frame is processed using a *Word2Vec* model from *Gensim* [6, 7], especially that built on *Google News Database* [8]. A *Word2Vec* model is Natural Language Processing (NLP) model in which a neural network is trained with a huge number of texts in order to assign to every word a vector representation that should describe its semantics. With the choice of *Word2Vec* model trained on *Google News Database*, we were able to cover approximately 3 million words. Using this model we get a 300-dimensional representation of the semantics of a single word. In this vector space the semantic similarity is highly correlated with the cosine distance between word vectors, since the elements of vectors can assume values in the range $[0, 1]$.

The word vector model consists of a matrix of m word vectors, for each word in the articles, as rows. On the other hand each column represents an n-dimensional feature vector, with n = 300.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \tag{1}$$

Then, we used a $tf \times idf$ term to weight the words in the articles, this is justified by natural language analysis [9]. This *term frequency-inverse document frequency* matrix is able to quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus).

Using the weights given by the $tf \times idf$ matrix is possible to exploit a weighted sum of the words which compose the article. We were able to compute the 300-dimension vector that describes the semantics of the entire article, i.e. for a document d containing k distinct words, its vector representation $\vec{D}$ is given by:

$$\vec{D} = \sum_{i=1}^{k} \omega_i \times W_{\omega,d} \tag{2}$$

where $W_{\omega,d}$ is the weight of the word ω_i in the document d.

Document vectors can be represented as the following matrix:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & \dots & y_{1n} \\ y_{21} & y_{22} & y_{23} & \dots & y_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{a1} & y_{a2} & y_{a3} & \dots & y_{an} \end{bmatrix}, \tag{3}$$

where each row represents the entire vector representation of the article, a represents the total number of articles taken into account.

The similarity between two articles is then defined as the cosine of the angle between two multidimensional vectors of two different article in this 300-dimensional space.

$$S = \cos\theta = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{k} A_i B_i}{\sqrt{\sum_{i=1}^{k} A_i^2}\sqrt{\sum_{i=1}^{k} B_i^2}} \tag{4}$$

For each couple of articles A,B and angle θ between their vector representations, i.e. $k = 300$.

3.2 Time window

The similarity between articles is not the only parameter to be considered for the epidemic model. It is, in fact, unreasonable that one article can *infect* a second one if it is published after the latter.

Moreover, following the same heuristic way to describe the model, further considerations has to be taken into account to treat the dynamic evolution as an epidemic process. These considerations, referring to the time window analysis, are the following.

- A minimum time threshold: as one may think, the articles could be infected only if the journalist is able to read the news from other sources, and this lead to a minimum time threshold, i.e. could not last minutes for an infection to happen, but hours.
- A maximum time threshold: this is an upper limit in time for one article to be considered influenced by a previous one, i.e. it is unreasonable to consider an article to be infected by another one published weeks before.

The calculation of the time window in which the model is applicable passes from building a time matrix, that takes into account the publishing date of each single article in the data set.

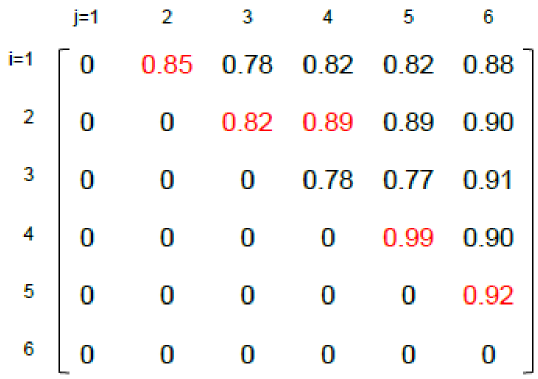
Using the time matrix it is possible to study the best time window for the model. Regarding the minimum time threshold it has been fixed, empirically, at 4 hours. This value takes into account what is the minimum time needed for one article to be read by a journalist and to influence the journalist to write about the same topic. For what concerns the choice of the maximum time threshold, we had to analyze the fraction of articles in the giant component of the reconstructed network for different values of this threshold: the value of the maximum time window was chosen to get, in the giant component of the reconstructed network, at least 80% of the total articles. This is a sort of constraint that the authors of the model fixed for building the reconstructed network, and it will be clearly shown in section 4.1

3.3 Similarity Matrix

The architecture of the network must be defined in order to reconstruct the spread of the news: the nodes represent the articles published on the subject chosen, and the edges are the infection events. The key principle to build such a network is: for every article after the first selected by the query, it must have been influenced (infected) by a previously published article.

Each article (except the first one) has been influenced by another, but its infector must fall into the *time window* defined in the section above. Following this concept the similarity matrix is an upper triangular matrix, meaning that only the articles already published can influence the succeeding ones and the previous article with maximum similarity is considered the infector.

The similarity matrix can be represented as in Fig. 2, each article in the j-th column shows a similarity coefficient (defined in Eq. 4) only for the articles published inside the time window and before its publication. Moreover for each article j its infector i has been highlighted in red and defined searching for the $max(s_j)$, where s_j are the similarity coefficients of Fig. 2.



$$\begin{array}{c|cccccc|}
 & j=1 & 2 & 3 & 4 & 5 & 6 \\
i=1 & 0 & 0.85 & 0.78 & 0.82 & 0.82 & 0.88 \\
2 & 0 & 0 & 0.82 & 0.89 & 0.89 & 0.90 \\
3 & 0 & 0 & 0 & 0.78 & 0.77 & 0.91 \\
4 & 0 & 0 & 0 & 0 & 0.99 & 0.90 \\
5 & 0 & 0 & 0 & 0 & 0 & 0.92 \\
6 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Figure 2: Example of the similarity matrix, where the red coefficients represents the maximum similarity for each column.[1]

3.4 Simulation

As previously said, the dynamic of the news spread is modelled using a SIR model for epidemic using the formalism of Pastor-Satorras el al. [2]. Whereas the implementation of the simulation follows the one proposed from Mussumeci and Coelho in [1].

Such simulation assumes that instead of modeling the status of a given individual as Susceptible (S), Infectious (I) or Recovered (R), one has to model the probability of each article being in each of the states. In this case, an article in status **S** would be one which has yet to be published, an article in status **I** one which is published and has been infected by the story and an article in status **R** is one which is too old to influence new articles. The SIR model leads to the system of Eq. 5 where $\rho_i^I(t)$ represents the probability of the article i to be infected at time t, $\rho_i^S(t)$ is the probability to be susceptible at time t, and λ is the adimensional transmission parameter which will be studied in the simulation process to match with the real data.

$$\begin{cases} \frac{d\rho_i^I}{dt} = -\rho_i^I(t) + \lambda\rho_i^S(t)\sum_{j=1}^N a_{ij}\rho_j^I(t) \\ \frac{d\rho_i^S}{dt} = -\lambda\rho_i^S(t)\sum_{j=1}^N a_{ij}\rho_j^I(t) \end{cases} \tag{5}$$

In the end a_{ij} represent instead the generic element of the adjacency matrix of the network A, which tells the probability of article j to be influcenced by article i. The adjacency matrix can be represented as:

$$\begin{cases} i = j : a_{ij} = 0 \\ i \neq j : a_{ij} = \frac{N_{IJ}}{N_J} \end{cases} \tag{6}$$

where the $\frac{N_{IJ}}{N_J}$ is an empirical influence fraction given by: N_{IJ} which is the number of article published from J (the publisher of article j) which have been infected by the publisher I (the publisher of article i) and N_J is the total number of articles of publisher J which has been infected from all the articles in the dataset.

From the solutions of the system of equations 5 we get the temporal evolution of the probabilities $\rho_i^I(t)$ and $\rho_i^S(t)$. Taken the solutions we were able to derive the realizations of the states for each article at time t $S_i(t)$, $I_i(t)$ and $R_i(t)$. Sampling from the probability distribution of states at each time step t, conditioning on the previous state, one has to follow the following procedure to reconstruct the states from t=0 until the final step:

- The binary state vectors for the articles at time t are S_t, I_t and R_t. Where 1 means that the articles belongs to that state in that time;
- For each $t > 0$ generate a newly vector I_t^* to represent the infected at that time step as a realization of a Bernoulli event with probability $\rho_i^I(t) \times S_{t-1}[i]$;
- As for the for the previous step generate a new vector R_t to represent the recovered at that time step as a realization of a Bernoulli event with probability $\rho_i^R(t) \times I_{t-1}[i]$;
- Update $I_t = I_{t-1} - R_t + I_t^*$ and $S_t = S_{t-1} - I_t^*$

Doing so we are able to get the state matrix I, which represents the infected articles at time t by their binary state 0 or 1. To describe the spread network we have to find which are the infectors for each time. Recalling the probability matrix A (Eq. 6), we have to iterate from $t = 1$ up to the final step and compute the probable infectors P_i as follows:

- For each article i in an infected state at time t multiply I_{t-1} by the $j - th$ column of matrix A, where $j = i$ to obtain its probable infectors P_i;
- Find the infector of article i by sampling from a multinomial distribution with $p = P_i$

Here a graphical example of the simulated spread taken from [1]:

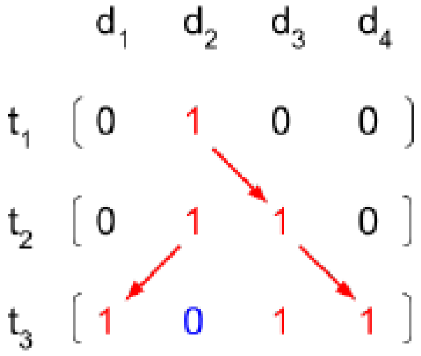


Figure 3: The arrows indicates the infector for each article. The red articles are the ones infected, that is, the ones that can spread the infection, and the blues are the ones that had recovered.

4 Results

Here the results discussed belong to the Norway summer school attack story, but the same concepts are general and are applicable to the other news. As it was described at the beginning, the construction of the network strongly depends on the selection of threshold parameters.

In order to understand how this dependency works, a 3D graph, shown in Fig. 4 has been studied. This represents a setup for selecting *hyper parameters* usually called "grid search". Here we performed a variation w.r.t. the original selection procedure, since authors of [1] did not use this specific grid search for the threshold *hyper parameters*. This "grid search" should be sufficient to select the *hyper parameters*, however we propose in the following the same selection method used in the paper.

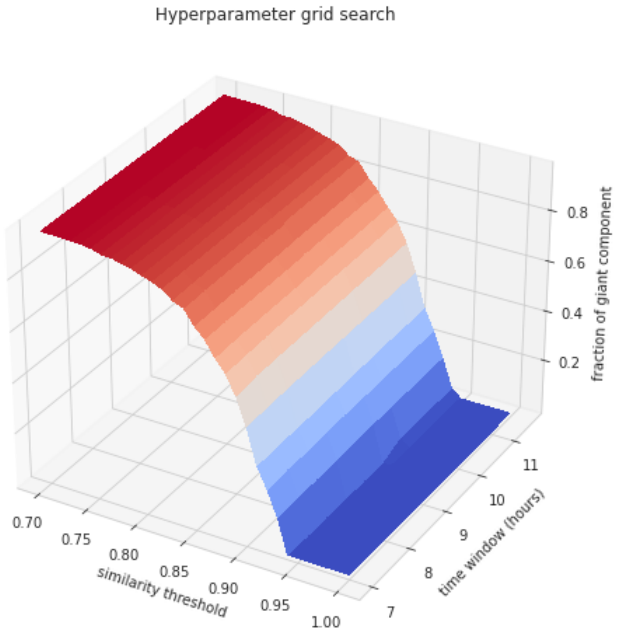


Figure 4: Dependency of the giant component by the time and the similarity

Looking at Fig. 4 it is clear that the fraction of articles in the giant component depends on the time threshold but even on the similarity threshold chosen. The similarity threshold arise since the articles belong to the result of a very specific search on a news articles database, therefore we can expect the articles to display a great similarity among themselves.

4.1 Selection of the maximum time threshold for the time window

First of all we have to define the time window γ for the articles to be influenced/influencer. The minimum time threshold between two articles, as previously displayed in section 3.2, has been set empirically. In fact as it is shown in Fig. 5 the distance in time between two consecutive articles in the data set is low.

The choice of a minimum time threshold of 4 hours is then useful to avoid that two consecutive articles, which are published almost at the same time, could be considered one influenced by the other. On the other hand for the maximum time threshold is fundamental to look at the time distribution of the articles. In Fig. 6 is represented the distribution of distances in time w.r.t. the

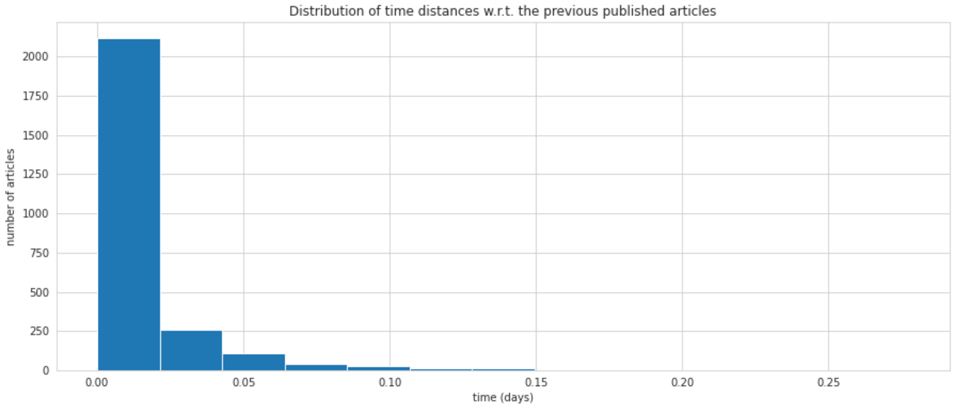


Figure 5: Distribution of time distances between one article and the previous one.

first published article.

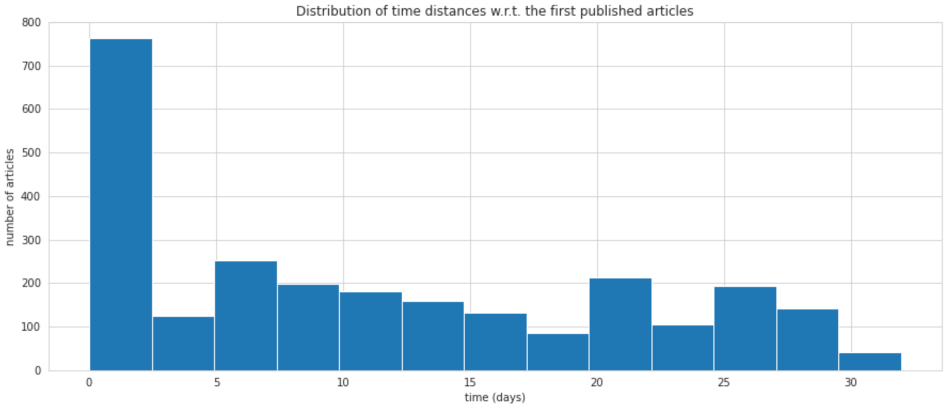


Figure 6: Distribution of articles published per day.

We were then able to determine the maximum threshold for the time window γ looking at the distribution of time lags from the infuencer article, that is shown in Fig. 7.

It is clear from the red dashed line that taking a maximum time threshold of 16 days we are able to cover the 95% of the most similar pairs in our articles data set. Then the time window $\gamma = 212$hours, taking in to consideration that $\gamma_{min} = 4$hours and $\gamma_{max} = 216$hours.

4.2 Selection of the similarity threshold

After the selection of the time window, the similarity has to be taken into account. The best way to study how the similarity affects the fraction of the giant component is to study it separately from the time, and then take the time window γ into consideration when it comes to the selection of the

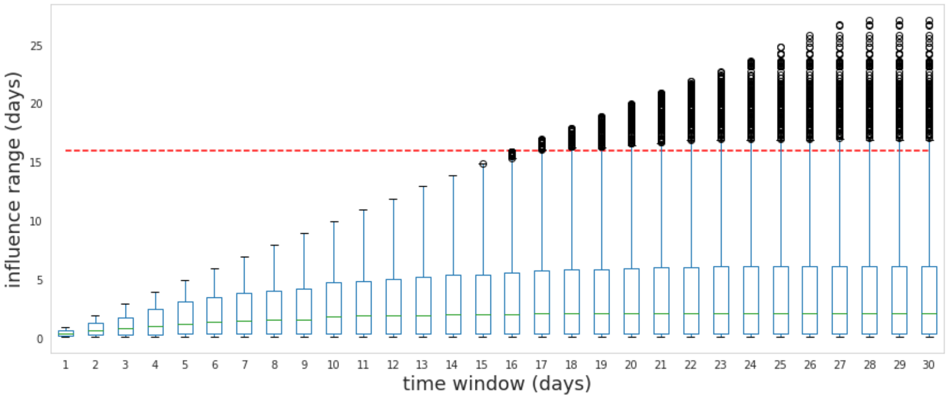


Figure 7: Boxplot representing the distribution of time lags from infuencer for different time windows from 1 to 30 days (the range selected in querying for articles). Notice that no article lags more than 9 days from its infuencer.

threshold for the giant component.

In Fig. 8 is represented the distribution of the pairwise similarity between all the articles of the data set.

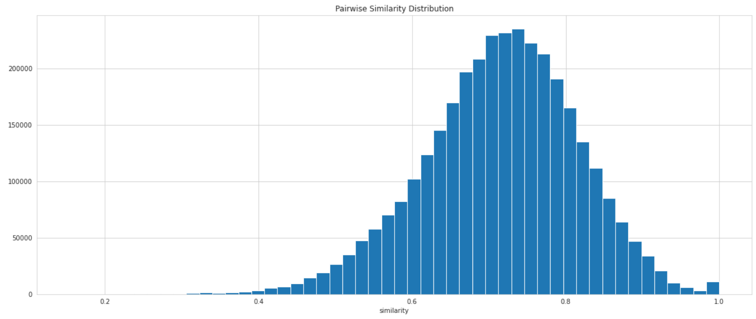


Figure 8: Distribution of the pairwise similarities among all articles.

Since the articles describe the same story, their similarity calculated in the 300 dimensional space is very high. This concept is even more clear if we plot the distribution of the most similar pair for each article (Fig. 9).

We can notice that for almost every article there is at least one other with similarity equal or greater than 0.8. We then built the reconstructed network using different values for the similarity threshold and we got the fraction of the giant component of the reconstructed network. Results are plotted in Fig. 10 showing how different values of the similarity threshold modify the fraction of giant component.

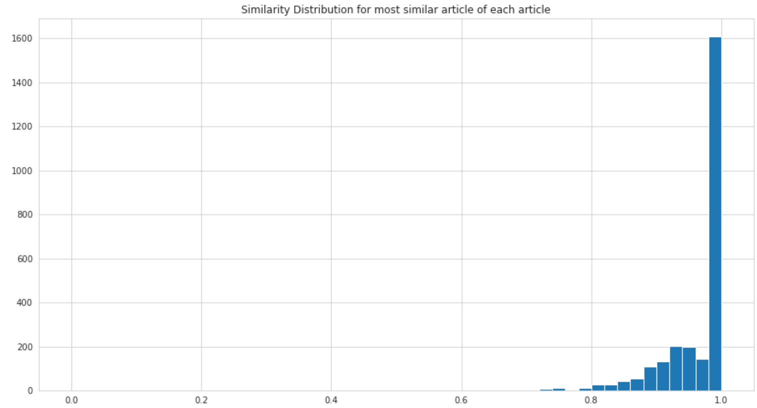


Figure 9: Distribution of similarities of the most similar article to each article in the collection.

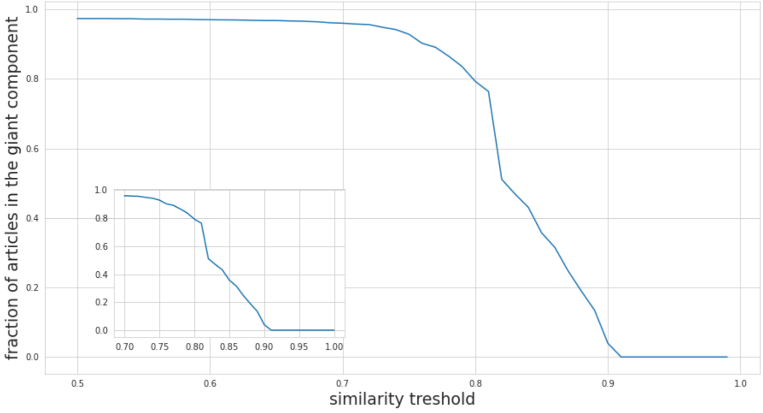


Figure 10: Fraction of articles in the giant component vs Similarity threshold.

In order to have a giant component in the network that contains at least 80% of our articles, we need to consider a minimum similarity of 0.8. Therefore, we define the similarity threshold $\rho = 0.8$. At the end, it is possible to graphically represent the network evolution by arranging them by generation, as shown in Fig. 11

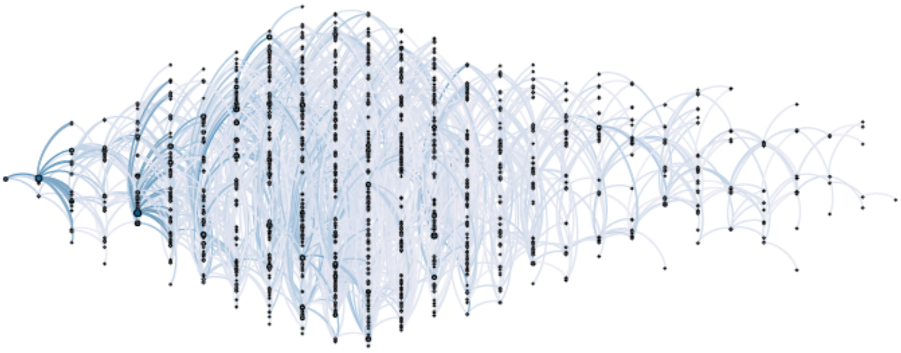


Figure 11: Network representation where the first column is the node that was first published, the second one are those influenced by him and so on. Taken from [1].

4.3 Simulation

After setting the *hyper parameters* for the similarity threshold ρ and for the time window γ, the last step left, before to compare the experimental data with the simulation explained in section 3.4, is to choose the adimensional transmission parameter λ.

Looking at the Fig. 12 which represents the number of articles published regarding the Norwegian attack day by day, it can be possible to estimate the maximum number of articles published at the peak of the diffusion. This value is what will lead us to the choice of λ.

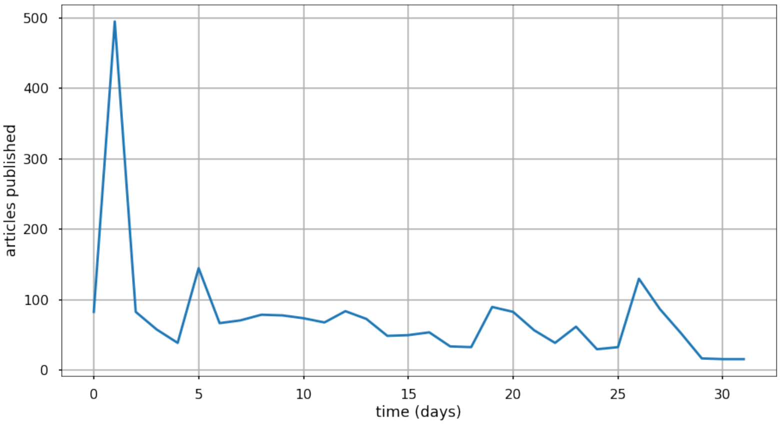


Figure 12: Number of articles published per day.

The value of the peak is of 500 articles. We then simulated the diffusion model for 100 different values of λ sampled from a log-uniform distribution in the range $[10^{-7}, 10^{-2}]$. We retrieved for every simulation the peak of infected in the population, Fig. 13 shows the results. In order to zoom this plot we performed 100 more simulation for values of λ sampled from a log-uniform distribution in the range $[10^{-4}, 10^{-3}]$, Fig. 14 shows the results. We were then able to set properly the value of λ to 1.5×10^{-4}.

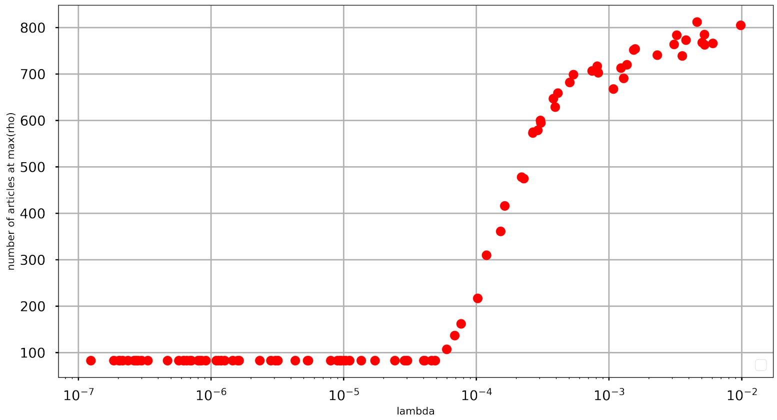


Figure 13: Total number of articles infected in the range of $10^{-7} < \lambda < 10^{-2}$

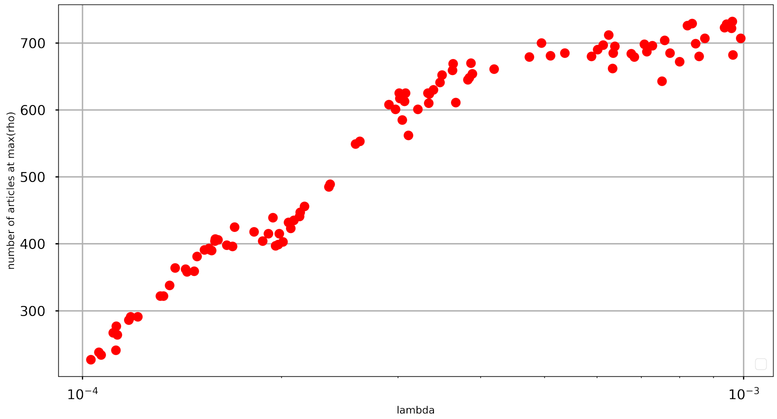


Figure 14: Specific simulation for $10^{-4} < \lambda < 10^{-3}$

At the end using the extracted value of λ we simulated 100 times the diffusion, results are plotted along with the real evolution of publications in Fig. 15.

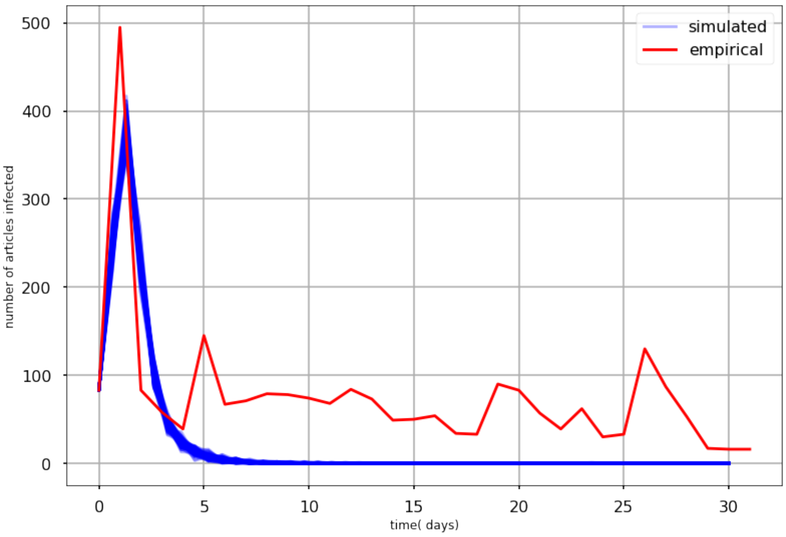


Figure 15: The blue curves are the 100 realizations of the state matrix. Notice that the simulated dynamics seems to have the same behaviour as the empirical curve

The entire code we used is available a public repository on *GitHub* [10].

5 Conclusions

The purpose of this project was to study the global spread of the news using a procedure to build a news spread network about a specific story and then study the diffusion of the story using an epidemic Susceptible-Infect-Recovered (SIR) model apply to that network.

The results proposed started from a well defined subset of articles with high semantic similarities, which may lead to the idea that this model is applicable only to similar articles. But as confirmed in [1] the similarity criteria used to reconstruct the network and the study of the *hyper parameters* can be used to characterize the news spreading even in generic groups of articles which do not share any common topic, given a large enough number of data.

The generic SIR model used for this study seems to approximate very well the empirical behaviour, at least for the first days. This could be explained as a typical shape for the specific event considered (the Norway attack) or it can be further analyzed with some changes in the model. Kalman filters [11] could be applied to correct the prediction for real time studies. Another improvement should be to apply other epidemic models, e.g. those that take into account vaccinated, to study the dynamics of fake news spread, in order to represent a certain degree of resistance to those kind of news.

In the end, the relevant result reached by this project is the following: it is possible to estimate with a good approximation the overall peak and time of persistence in the media of a story by observing the very beginning of news spreading, this could be done exploiting the application of a SIR model to a properly constructed network of articles.

References

[1] E. Mussumeci and F. Coelho, “Reconstructing news spread networks and studying its dynamics,” *Social Network Analysis and Mining*, vol. 8, 01 2018.

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, “Epidemic processes in complex networks,” *Rev. Mod. Phys.*, vol. 87, pp. 925–979, Aug 2015. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.87.925

[3] H. Roberts, R. Bhargava, L. Valiukas, D. Jen, M. M. Malik, C. Bishop, E. Ndulue, A. Dave, J. Clark, B. Etling, R. Faris, A. Shah, J. Rubinovitz, A. Hope, C. D’Ignazio, F. Bermejo, Y. Benkler, and E. Zuckerman, “Media cloud: Massive open source collection of global news on the open web,” *CoRR*, vol. abs/2104.03702, 2021. [Online]. Available: https://arxiv.org/abs/2104.03702

[4] Media Cloud developers, “Python api tutorials.” 2022. [Online]. Available: https://github.com/mediacloud/api-tutorial-notebooks

[5] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” in *IN IPROCEEDINGS OF THE 21ST NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE - VOLUME 1*, 2006, pp. 775–780.

[6] Radim Řehůřek. (2021) Gensim. [Online]. Available: https://radimrehurek.com/gensim/

[7] RARE Technologies. (2022) Rare technologies repository. [Online]. Available: https://github.com/RaRe-Technologies/gensim-data

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[9] D. Hiemstra, “A probabilistic justification for using tf-idf term weighting in information retrieval,” *Int. J. on Digital Libraries*, vol. 3, pp. 131–139, 08 2000.

[10] Lorenzo Sani and Daniele Buschi. (2022) Cn project. [Online]. Available: https://github.com/relogu/cn_project

[11] R. Lal, W. Huang, and Z. Li, “An application of the ensemble kalman filter in epidemiological modelling,” *PLOS ONE*, vol. 16, no. 8, pp. 1–25, 08 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0256227