TENDÊNCIAS ATUAIS E PERSPETIVAS FUTURAS EM ORGANIZAÇÃO DO CONHECIMENTO

ATAS DO III CONGRESSO ISKO ESPANHA-PORTUGAL
XIII CONGRESSO ISKO ESPANHA

Universidade de Coimbra, 23 e 24 de novembro de 2017

Com a coordenação de

Maria da Graça Simões, Maria Manuel Borges

TÍTULO
Tendências Atuais e Perspetivas Futuras em Organização do Conhecimento: atas do III Congresso ISKO Espanha e Portugal - XIII Congresso ISKO Espanha

COORDENADORES
Maria da Graça Simões
Maria Manuel Borges

EDIÇÃO
Universidade de Coimbra. Centro de Estudos Interdisciplinares do Século XX - CEIS20

ISBN
978-972-8627-75-1

ACESSO
https://purl.org/sci/atas/isko2017

OBRA PUBLICADA COM O APOIO DE

· U C ·

FLUC FACULDADE DE LETRAS
UNIVERSIDADE DE COIMBRA

C E I S 2 0
CENTRO DE ESTUDOS
INTERDISCIPLINARES
DO SÉCULO XX
UNIVERSIDADE DE COIMBRA

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

PROJETO UID/HIS/00460/2013

RECONSTRUCTING NEWS SPREAD NETWORKS AND STUDYING ITS DYNAMICS

Elisa Mussumeci[1], Flávio Codeço Coelho[2]

[1]*Fundação Getulio Vargas, 0000-0002-7882-6465, elisamusumeci@gmail.com*
[2]*Fundação Getulio Vargas, 0000-0003-0635-8989, fccoelho@fgv.br*

ABSTRACT News spread can be seen as a contagious process in internet media outlets. This process can be transformed into a temporal network which represents the influence between published articles and between media outlets. In this article, we propose a methodology based on the application of natural language analysis of the articles to reconstruct the latent network through which news spread. From the reconstructed network, we analyze the network dynamic and then show that the dynamics of the news spread can be approximated by a classical SIR epidemiological dynamic upon the network. From the results obtained we argue that the methodology proposed can be used to make predictions about media repercussion, and also to detect viral news in news streams.

KEYWORDS *News, SIR, Epidemics, Temporal Networks*

INTRODUCTION

The Internet is the main channel for information dissemination in the 21st century. Information of any kind is posted online and spreads via recommendations or advertisement (Hermida et al. (2012), Romero et al. (2011)).

Studying the spread dynamics of information through the internet, is a very relevant and challenging activity, since it can help the understanding of the factors which determine how far and how fast a given information can go. The most common way to observe information flow on the web, is by tracking how many times a given piece is replicated by different users over a period of time. Sometimes the content is modified as it is replicated, making it harder to track. This type of dynamics is common to news articles, social media posts, and a wide variety of content which gets replicated in digital networks.

In the specific case of news articles, a number of factors influence their spread. Among the most important are the reputation of the original publisher -- though that is not easy to measure, and the size of readership of a particular publisher, which will determine the initial spread of any piece. However, the topology of the resulting network associated with dissemination of news is hard to predict and will depend on the subject of each news piece and its resonance with public interests.

In this work, we decided to look at the spread of news stories over the Internet characterizing the resulting spread network and the dynamics of the spread. We start by looking at an actual case of news spread, and reconstructing the spread network by applying ideas of temporal networks and topic modeling, connecting similar articles within the bounds of temporal window of influence. Then we postulate that the spread dynamics approximates an epidemic process and model it using a Network SIR model (Pastor-Satorras et al. (2015)). Modeling the spread of ideas as an epidemic process is not a new idea (Bettencourt et al. (2006)), but here we propose new tools to estimate the spread network from data and compare it with simulated networks produced by an SIR epidemic model. By studying the topologies of these networks and the typical dynamics, we believe it is possible to detect anomalies in the propagation of news in digital media.

METHODOLOGY

DATA SOURCES

The data used for this study was obtained from the Media Cloud Brasil project (MCB) which collects news articles from thousands of sources in the Brazilian Internet since 2013. From the MCB database we obtained 2129 articles talking about the Charlie Hebdo terrorist attack in February 2015. The corpus of articles spans from the day of the attack to the end of march of 2015. The data include the full text of the article, the URL of publication and the date and time of the publication. The choice of topic was guided by the need to have a well defined news topic that was not very likely to be confounded with other stories being published in the same period.

ARTICLE SIMILARITY

To calculate a measure of similarity between text documents one can rely on a number of metrics for textual distance described in the literature (Mihalcea et al. (2006)). Most of these metrics are based on a bag-of-words representation of texts, meaning that texts are defined in terms of which words they contain and their respective frequencies in the documents. Such representation completely disregards higher level linguistic features of texts such as syntactic and semantics. In this analysis, we want to use semantically similarity to describe the association between articles. In order for a news article to influence another, they must talk about the same concepts.

In order to capture the semantics of the articles we started by building a word vector representation for every word in our corpus' vocabulary, taking into account the co-occurrence of words within a sentence. This model is built from a larger corpus of news articles (approximately 2.6 million articles) according to the Skip-gram model, which has been shown to map the words to a vector space where semantic similarity is highly correlated with the cosine distance between word vectors (Mikolov et al. (2013)). This larger corpus corresponded to the total collection of articles of the MCB project at the time this analysis was done. The importance of training the word vector model on a corpus as large as possible, is that one gets a more accurate semantic representation of each word as a vector. It is important that the larger corpus represents a similar informational space as the sample we are trying to analyze.

Table 1. Parameters of the skip-gram model used to construct the word vectors.

Parameter	Value	Meaning
Minimum word count	10	Word minimum frequency in the corpus
Number of features	300	Dimension of word vectors
Context	10	Text window around word

The word vector model, was trained with the parameters described in table 1. The fitted word vector model consists of a matrix of m word vectors (w_i) as rows. Each row represents a n-dimensional feature vector, with $n = 300$:

$$\begin{array}{cc} Pm & \begin{array}{cccc} f_1 & f_2 & \dots & f_n \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ \vdots \\ w_m \end{array} & \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

From the word vectors obtained, we created document vectors defined as a weighted sum of word vectors. For a document d containing k distinct words, its vector representation D is given by:

$$\vec{D} = \sum_{i=1}^{k} w_i \times \mathcal{W}_{w,d} \tag{1}$$

where $w_{w,d}$ is the weight of the word w_i in the document d. This weight can be calculated in different ways, for this work we used the TFIDF score (Hiemstra (2000)) of the word in the document. Another possibility would be to use the frequency of the word in the document.

The TFIDF weight is the product of two numerical statistics: term frequency and inverse document frequency. There are several ways of calculating this weight, the one used in this article was the below:

$$f_{ij} * \log\left(\frac{D}{df_i}\right),$$

where f_{ij} is the frequency of the word i in document j, D is the corpus, and df_i is the number of documents where the term i appears. The second term of this equation is the IDF, that measures how much information the word provides, that is, it's shows us if it's a common word or a rare one.

From the weighted sum, we obtain document vectors which can also be represented by a matrix:

$$\begin{array}{cc} & \begin{array}{cccc} f_1 & f_2 & \dots & f_n \end{array} \\ \begin{array}{c} d_1 \\ d_2 \\ \vdots \\ d_M \end{array} & \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{Mn} \end{bmatrix} \end{array}$$

Now we can define the similarity between two documents $\{A, B\}$ as the cosine of the angle θ between their vector representations:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}} \qquad (2)$$

TEMPORAL ASSOCIATION

Once the similarity of two articles is calculated, their temporal association must be determined in order to consider the probability of the older article being the *infector* or influencer of the more recent one. To determine the most-likely *infector* of an article, we ranked all articles by date of publications and looked within a fixed time window preceding the publication of each article, for the articles which are most semantically similar. The choice of the size of the time window was determined in order encompass the majority ($\geq$ 95%) of previous similar articles (see figure 4).

RECONSTRUCTING THE SPREAD NETWORK

To reconstruct the spread network of the news, we defined the nodes of our network as the articles published on the subject chosen, and the edges as the infection events, i.e., for every article after the first one, it must have been influenced (infected) by a previously published article. To qualify as an

infector, an article must precede the infected article by less than γ hours, and have a score of similarity (defined by equation (2)) to the infected article of at least ρ. The reconstruction procedure is summarized in the four steps below.

1) Rank all articles in ascending publication time. Let p_i denote the publication date of article i.
2) Create upper triangular matrix D, where d_{ij}=H($\gamma\delta_{ij}$)\*δ_{ij} and δ_{ij}=p_j-p_i. H is the Heaviside function.
3) Create similarity matrix S. Where s_{ij} is the similarity defined by equation (2) whenever $d_{ij} \neq$ 0 and 0 otherwise.
4) For each article j, we define its influencer i as the article corresponding to $max(s_j)$.

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

Figure 1: This figure is a slice of similarity matrix *S*, it shows the first 6 articles. In red are the maximum similarity score for each column, which we use to define it's infector, per example, the article 4 has been infected by the article 2.

SIMULATION MODEL

To test the hypothesis that news spread follows an epidemic process, we proposed an SIR model for the spread, following the formalism of (Pastor-Satorras et al. (2015)). In this formalism, instead of modeling the status of a given individual as Susceptible (S), Infectious (I) or Recovered (R), we model the probability of each article being in each of the states, in this case, an **S** article would be one which has yet to be published, and **I** one which is published and has been infected by the story and an **R** is one which is too old to influence new articles. This modeling leads us to equations (3).

$$\begin{cases} \frac{d\rho_i^I}{dt} & = -\rho_i^I(t) + \lambda\rho_i^S(t)\sum_{j=1}^N a_{ij}\rho_j^I(t) \\ \\ \frac{d\rho_i^S}{dt} & = -\lambda\rho_i^S(t)\sum_{j=1}^N a_{ij}\rho_j^I(t) \end{cases} \qquad (3)$$

In equation (3), $\rho_i^I(t)$ is the probability of article *i* being in the infectious state at time *t*, similarly for $\rho_i^S(t)$; a_{ij} is the probability of article *j* being influenced by *i* and comes from the adjacency matrix of the network. λ is an adimensional transmission parameter given by $\lambda = \frac{\beta}{\mu}$. Time (*t*) in these equations is also adimensional as it is scaled by μ.

The network for the simulation is built from the same node set of the empirical data. The adjacency matrix *A* is given by

where N_{XY} is the number of times an article from publisher *X* (the publisher of article *i*), has infected an article from publisher *Y* (the publisher of article *j*) and N_Y Is the total number of articles from publisher *Y* that have been infected, regardless of publisher. These counts are derived from the empirical dataset.

$$\begin{cases} i = j : a_{ij} = 0 \\ i \neq j : a_{ij} = \frac{N_{XY}}{N_Y} \end{cases} \qquad (4)$$

The solution of this model generates the temporal dynamics of the probabilities described in (3). From the solutions, $\rho_i^S(t)$ and $\rho_i^I(t)$ we can derive realizations of states for each article, $S_i(t)$, $I_i(t)$, and $R_i(t)$.

To reconstruct the states, we must sample from the probability distribution the states at each time *t*, conditioning on the previous state. We follow the procedure:

1) Let S_t, I_t and R_t be binary state vectors from article states at time *t*, where 1 means the article is in that state.
2) Iterate from $t = 0$ until the final time step available.
3) For each time $t > 0$ generate a newly infected I_t^* vector, in which each element i is a realization of a Bernoulli event with probability given by $\rho_i^I(t) \times S_{t-1}[i]$.
4) Similarly to the previous step, sample a new R_t vector, in which each element i is a realization of a Bernoulli event with probability given by $\rho_i^R(t) \times I_{t-1}[i]$.
5) Update $I_t = I_{t-1} - R_t + I_t^*$.
6) Update $S_t = S_{t-1} - I_t^*$.

At the end of the procedure above we'll have tree matrixes of states: *I, S* and *R*. Each matrix is composed by the states vectors (S_t, I_t, R_t) and represents the states dynamics of the model.

CONSTRUCTING THE SIMULATED SPREAD NETWORK

From the state matrix *I* we have which articles get infected at each time *t*. To create a spread network for the simulation we need to define the infectors for each time. For that we used the probability matrix *A* defined by equation (4). The following steps describe the entire procedure.

1) Let I_t be binary state vectors for articles at time t, where 1 means the article is infected.
2) Iterate from $t = 1$ until the final time step available.
3) For each article *i* infected at I_t, obtain its probable infectors, P_i, by multiplying I_{t-1} by the column *k* of matrix *A*, where $k = i$ and the values are the probability of each article *j* from I_{t-1} has to infect *i* (a_{ij} of the matrix A).
4) Define the infector of *i* by sampling from a multinomial distribution with $p = P_i$.

The figure 2 shows the procedure, where the rows t_i are the states vector and are composed by the state value for each document in the step *i:*

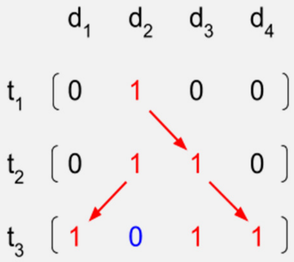

Figure 2: The arrows indicates the infector for each article. The red articles are the ones infected, that is, the ones that can spread the infection, and the blues are the ones that had recovered.

FINDINGS

The dataset used is the result of a very specific search on a news articles database, therefore we can expect to the articles to display a great similarity among themselves.

Looking to the distribution of pairwise similarities that were used to construct the empirical influence network, we can notice that for almost every article there is at least one other with similarity equal or greater than 0.8. Identical articles (similarity equals to 1) were not considered for edge formation.

Figure 3 shows the similarity threshold of the influence network. In order to have a giant component in the network that contains at least 80% of the articles, we need to consider a minimum of 0.8 score similarity. Therefore, we defined $\rho = 0.8$.

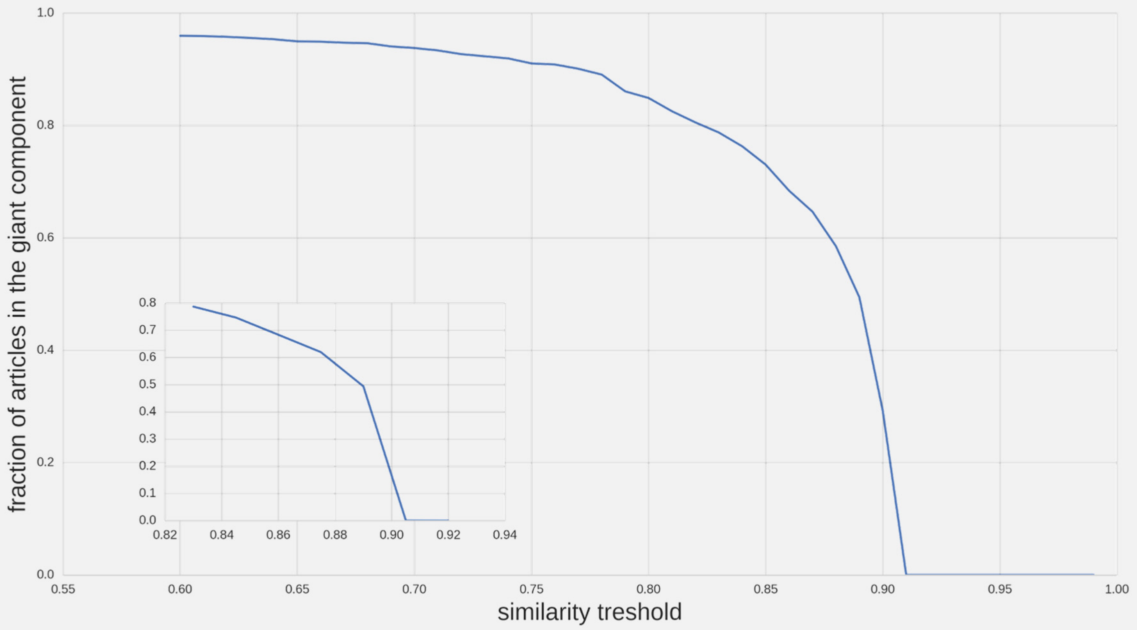

Figure 3: Similarity threshold for the reconstruction of the influence network.

To determine the optimal time window γ in which to search for influencers, we looked at the distribution of time lags from the most similar article (most likely influencer) at various window lengths (figure 4). Even for time windows as long as 15 days, 95% of the influencers where within 7 days of the articles they influenced.

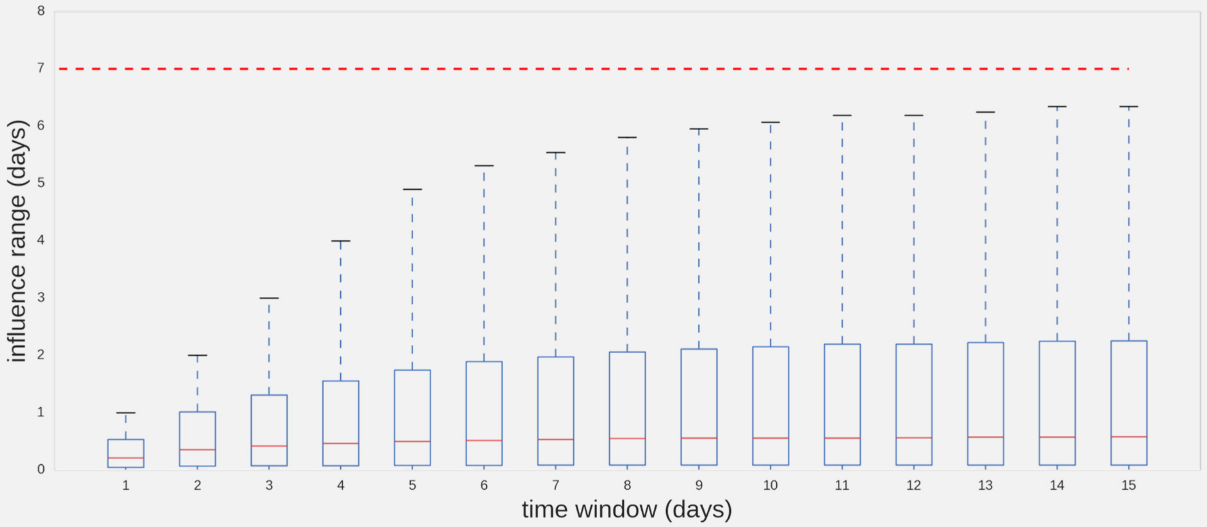

Figure 4: Distribution of time lag from influencer for multiple time window lengths. Notice that no article lags more than seven days from its influencer.

To create the spread network, we defined influence based on the time lag from each pair of articles and from their similarity. Following the previous analysis, we defined $\rho = 0.8$ and γ=168, which means the infector must precede the infected article by less than 168 hours (7 days) and have at least a 0.8 score of similarity.

Looking at the publication date distribution we notice that the maximum number of articles published in a day was between 250 and 300. We derive the simulation parameters from this distribution. For example, on figure 5 we plot the peak of the infection for a range ([0,0.00005]) of λ values. From that distribution of peak magnitudes we selected a λ to match the empirical peak: $0.00002 < \lambda < 0.00003$.

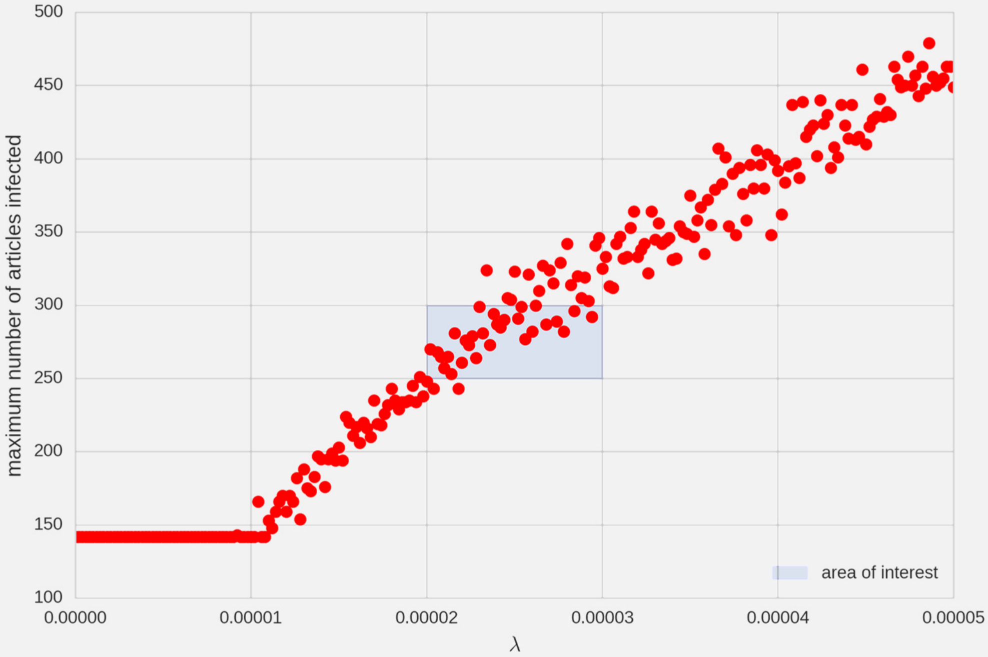

Figure 5: Total number of articles infected between $0 < \lambda < 0.00005$. The blue area is the area where the peak of the simulation is the same as the peak of the dataset distribution, therefore is the area where the λ values were tested for our simulation.

From the simulation (figure 6) we obtain the state matrix, which we use to compare the simulated infection distribution with the original data. Then we ran 10 thousand simulations to show that the model proposed matches the real world dynamics of news articles influences (figure 7).

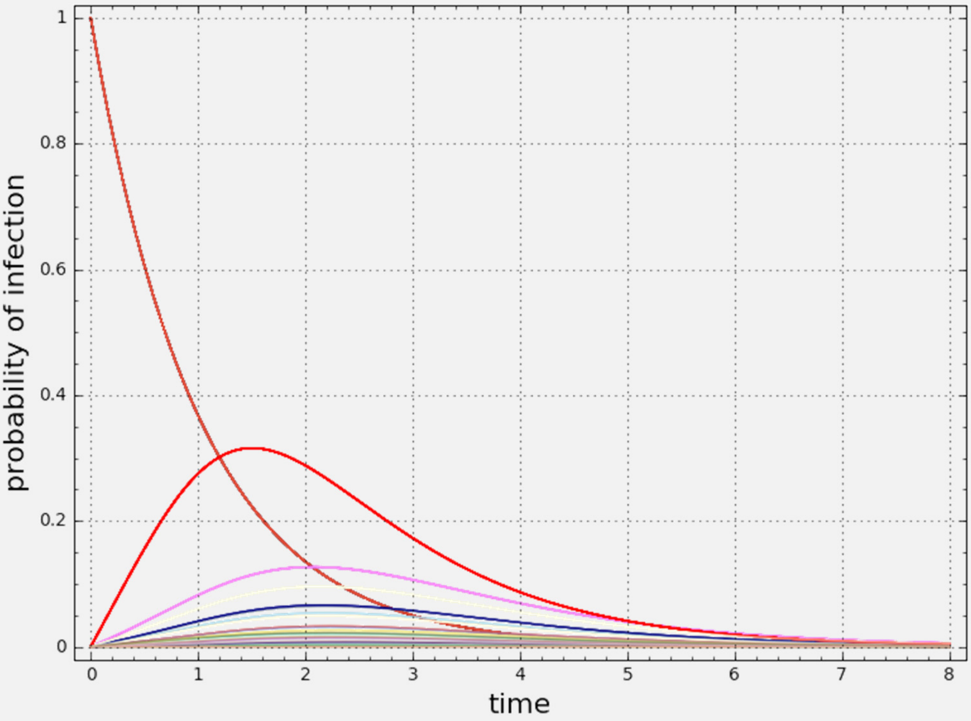

Figure 6: Simulation for $\boldsymbol{\lambda} = 0.0000215$. Each curve represents the n the infectious state as a function of time, for every article. The time units are $1/\boldsymbol{\mu}$.

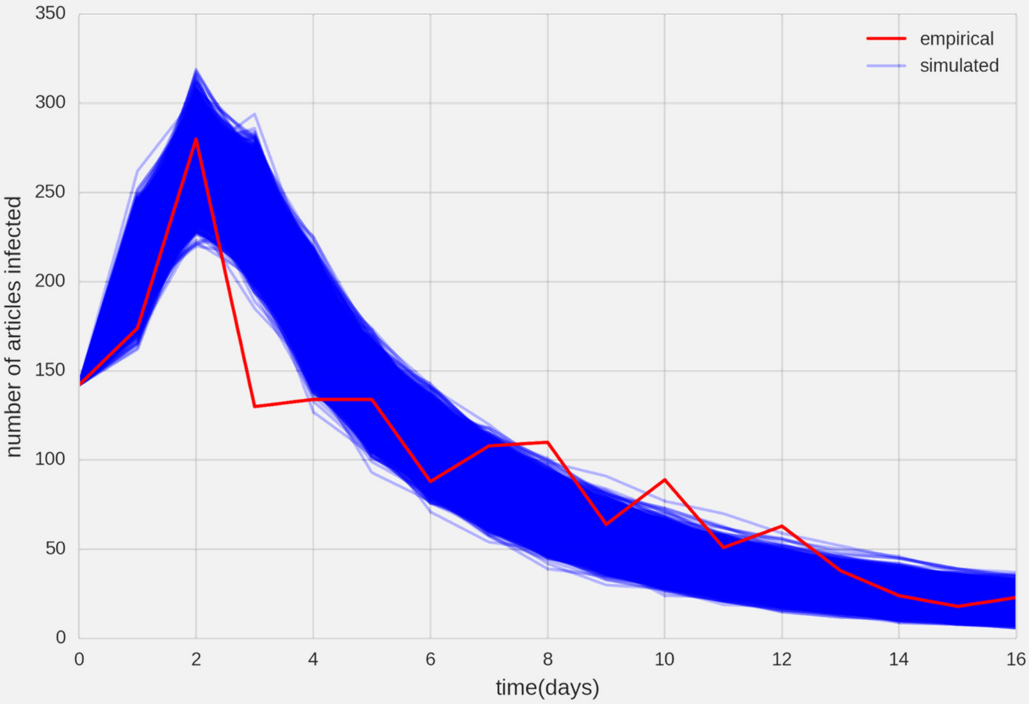

Figure 7: The blue curves are the 10.000 realizations of the state matrix. Notice that the simulated dynamics match the empirical curve.

CONCLUSIONS

In this paper, we presented a methodology for reconstructing the network representing the spread of news in digital media. The results proposed started from a well defined subset of articles with high semantic similarities. However, we believe the criteria of similarity used to reconstruct the network would work even on a random sample or articles, provided that it was large enough to contain a good portion of the putative spread network one is trying to characterize. In other words, the reconstruction algorithm can be used to detect contagious structures within any large enough collection of news articles. Although we have not tested, we believe that the article similarities could be obtained from a topic modeling algorithm less dependent on a large training corpus, such as the Latent Semantic Analysis model. We nevertheless still favors using the word embedding model presented here, as it is not that difficult to obtain reasonably large corpora for training such models. We remark that these corpus size requirements do not apply to the corpus of news whose network we want to reconstruct.

We also demonstrated that a classical SIR process over the network is driving the spread dynamics. This means that if one is able to observe the start of the spread, the overall reach and time of persistence in the media can be predicted from analytical results available for the SIR model, namely the early estimates of the basic reproductive number or R0 of the model, which can be calculated very early on from the number articles the first article influences.

The type of analysis introduced here could very easily be adapted to posts on social media, given that a reliable means to capture the full publication set is available. It is increasingly common to see mechanisms to boost social posts, by analyzing the spread dynamics of posts, could help discriminate boosted from unboosted posts. The dynamics of spread can also be used to measure the efficacy of said boosting mechanisms. More generally this methodology can be used for any kind of meme spread ranging from applications in marketing, to political propaganda. More research is needed to characterize the diversity of spreading patterns we can observe in digital media networks.

The news subject selected, "Charlie Hebdo attack", represents a very spontaneous media coverage given the great surprise with which it happened, but also due the homogeneous response of the global media condemning the cowardly attack. We believe that deviations from the classical SIR dynamics shown here can hint at some form of media manipulation, but that hypothesis remains to be tested based on well defined cases, such as purchased media coverage during political campaigns, etc. With the current concerns about "fake" media pieces (Berkowitz, Schwartz (2016)), Perhaps the methods presented here can help to discriminate authentic media articles from fake ones based on their spread dynamics or influence patterns. We already began to see some attempts to automatically detect fake news (Jin et al. (2016), Rubin et al. (2015)), but they mostly rely on linguistic cues. We believe that qualitative and quantitative aspects of the spread networks can also be of use, as they reveal media manipulation mechanisms, such as “pay to publish” or editorial biases of mainstream outlets which can block or favour certain themes.

REFERENCES

Berkowitz, D., & Schwartz, D. A. (2016). Miley, CNN and The Onion: When fake news becomes realer than real. *Journalism Practice*, 10(1), 1-17.

Bettencourt, L. M., Cintrón-Arias, A., Kaiser, D. I., & Castillo-Chávez, C. (2006). The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, 364, 513-536.

Hermida, A., Fletcher, F., Korell, D., & Logan, D. (2012). Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6), 815-824.

Hiemstra, D. (2000). A probabilistic justification for using tf×idf term weighting in information retrieval. International *Journal on Digital Libraries*, 3(2), 131-139.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia*.

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI* (Vol. 6, pp. 775-780).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3), 925.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, September). Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 18-33). Springer Berlin Heidelberg.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1), 1-4.