Università degli Studi di Catania

Dottorato di ricerca in
Ingegneria dei Sistemi, Energetica, Informatica e delle Telecomunicazioni – XXXII ciclo

Online Social Network News Spreading: Modeling and Data Analysis

Student: Marialaura Previti

Tutor: Prof. Ing. Vincenza Carchiolo

A.Y. 2018/2019

Issue

- The enormous use of online social networks (OSNs) in the dissemination of information has made these platforms famous for rumors spreading, but, unlike traditional channels, news is free to propagate without control, so it is of primary importance to understand what are the factors that induce users to propagate a piece of news and not another, and how to identify true and false news.
- The topic is of interest in multiple areas: psychology, philosophy, politics, marketing, business, finances, and so on.
- Social networks are constantly modifying the way users create, share and consume information, and were become powerful instruments for understanding social trends and the society behind them, so the purpose of Ph.D. work is dual.

Research activities

- In the first part, I talk about mechanism of social contagion and introduce **three different models based on credibility of posts and users who publish news**, showing the necessity of an intangible superstructure on the acquaintance network (naturally present in each social network) that take into account this parameter, in order to model the attitude of each user to propagate or not a piece of news.
- Considering the difficulty to obtain appropriate dataset to test the models, I also implemented three simulators in order to evaluate the influence of the credibility parameters.

Research activities

- In the second part, I analyze a dataset of Twitter data containing full cascades of tweets propagated between 2006 and 2016 and concerning specific topics:
- under the perspective of cascades, writing **a mathematical equation that approximate the retweet dynamic**;
- under the perspective of rumors, finding **a method to identify true and false news through the classification of the time series features and the information about users involved in news spreading**.

Model 1: Post-sharing based

- The decision to spread a piece of news depends on both the news credibility and the proposer credibility.
- My network is a duplex network composed by:
  - A **multislice acquaintance network**, where each slice represents the propagation of a piece of news;
  - A **directed credibility network**, where each edge represents the confidence of receiver of news respect to the spreader.

Carchiolo, V., Longheu, A., Malgeri, M., Mangioni, G., & Previti, M. (2017, October). Post sharing-based credibility network for social network. In *International Symposium on Intelligent and Distributed Computing* (pp. 149-158). Springer, Cham.

Model 2: Epidemic model for news spreading

- I use **SIR (Susceptible-Infected-Removed) model**, typically exploited to describe how disease spread over population, in social network context to describe how news spread over the network. With regard to each news, I divide social network users into 3 categories:
  - **ignorants**, i.e. people who are unaware about the news;
  - **spreaders**, i.e. people who are already aware about the news and intend to share it with others;
  - **stiflers**, i.e. people who are already aware about the news, but have no interest in spreading it.
- In my model, every time that a spreader node *i* decide to post or re-share a news, its ignorant neighbor node *j* use a **modified version of Newman's transmissibility formula Tij** (*Newman, M.E.: Spread of epidemic disease on networks*) to decide if re-post it or not.

$$T_{ij} = 1 - e^{-\beta_{ij}\tau_i Cr_{ji}}$$

- where *βij* is the contact rate between *i* and *j*, *τi* is the time interval in which node *i* is infectious, hence its news is visible (and sharable) for its neighborhood.

Model 2: Epidemic model for news spreading

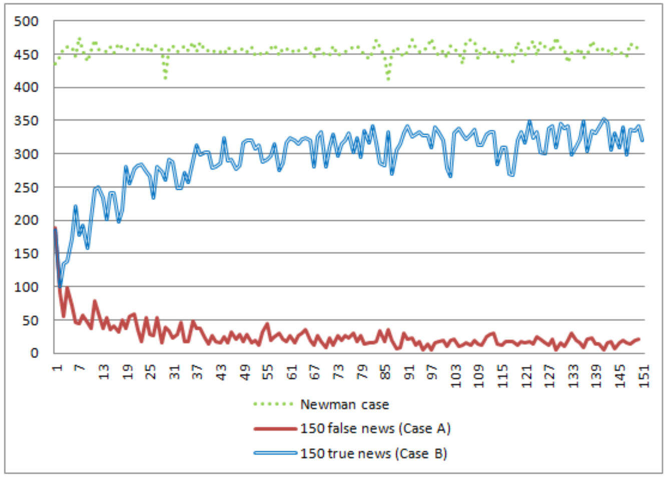
- To update credibility every time that a node *j* receives a news, we use an equation that perform a weight average of **news credibility Cnews**, taking into account to the aging of the news inserted by spreader *i* over time.

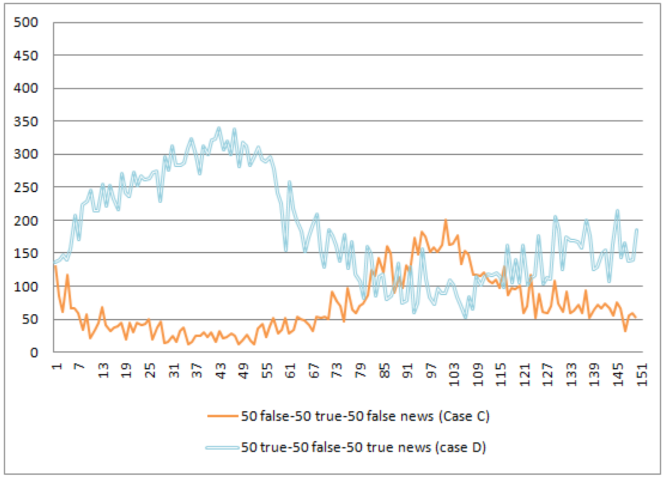
$$Cr_{ji} = \frac{\sum_{x=0}^{n}(x+1)C_{news_x}}{\sum_{x=0}^{n}(x+1)}$$

- where *Cnews* value range from 0, if the news *x* is patently false, to 1, if the news *x* is certainly true.

- Carchiolo, V., Longheu, A., Malgeri, M., Mangioni, G., & Previti, M. (2017, November). Introducing Credibility to Model News Spreading. In *International Conference on Complex Networks and their Applications* (pp. 980-988). Springer, Cham.
- Carchiolo, V., Longheu, A., Malgeri, M., Mangioni, G., & Previti, M. (2018, March). A trust-based news spreading model. In *International Workshop on Complex Networks* (pp. 303-310). Springer, Cham.

Model 2: Epidemic model for news spreading





Model 3: Trust and reliability for not competitive and competitive news spreading

- I used a duplex network composed by:
  - a **directed acquaintance network;**
  - a **directed weighted credibility network** with edges in opposite direction respect to the acquaintance network nodes, because, if a node spreads a piece of news, the receiving node forms an opinion about the spreader, modeled by the weight **$c \in [-1,1]$**.

Model 3: Trust and reliability for not competitive and competitive news spreading

- I embed two parameters in credibility network structure:
  - the **trust (T)** in an individual indicates how much he/she is considered trustworthy by its neighbors; high trust values indicate that who is in contact with him/her appreciates the contents he/she posted and considers him/her a person who verifies the news before reposting it (it is related to incoming edges of credibility network);

$$T_{t+1}(v) = \frac{1}{|v|_{in}} \sum_{u \in U_{in}(v)} R_t(u) c(u,v)$$

  - the **reliability (R)** of an individual shows its ability to select which neighbors he/she will accept news from to repost, hence this parameter indirectly influences his/her ability to post true news (it is related to outgoing edges of credibility network).

$$R_{t+1}(v) = 1 - \frac{1}{|v|_{out}} \sum_{u \in U_{out}(v)} \frac{|T_{t+1}(u) - c(v,u)|}{2}$$

Model 3: Trust and reliability for not competitive and competitive news spreading

- I took in consideration three kinds of news spreading models:
  - **not competitive news spreading model**, in which news is free to propagate over the network without opposition;
  - **competitive news spreading model**, in which there are two different thought factions about the same topic that propagate the news at the same time and a group of OSN users reached by both factions is called to decide which side to belong to;
  - **competitive news spreading model with delay on the second piece of news propagation**, in which there are two different thought factions about the same topic that propagate the news, but the second one starts after the first one propagation as denial of it and a group of OSN users is called to decide if, after the propagation of first news, publish the retraction.
- After the setting of the initial values on credibility network, I suppose that a set of nodes **S** becomes **spreaders**, activating themselves to propagate a piece of news. The **ignorant** neighbors **I** who are exposed to a piece of news must decide if to repost it or not and, to do this, the sum of the credibility of the edges linking with spreader nodes must exceed the activation threshold of the inactive nodes.
- In competitive cases, both the factions credibility are evaluated and nodes propagate the most influential one, but in delayed case the second piece of news is evaluated after the end of the first propagation, hence can be some nodes that post both pieces of news.

Data analysis: Retweets dynamic

- Observing the number of total tweets per cascade in 3 different cases (true, false and mixed news), I see that, despite the number of retweets is distinctly different, all the curves are monotonic increasing with some steps in correspondence with the peaks present in the tweet per hour graphs. So it is possible to identify a parameterized function that approximates the value of each curve.
- I generated a model starting from the observation of my data subset characteristics.
- In the following, I indicate with R(t) the number of retweets at instant t. The number of retweets in a generic instant t +Δt will be given by the sum between the ones at instant t and the ones that will be posted in the following instants with a certain retweet rate λ, but not exceeding the number of total retweets of the curve N, hence:

$$R(t+\Delta t) = R(t) + \lambda[N - R(t)]\Delta t$$

- In the limit of a small dt, I have the following first-order ordinary differential equation:

$$R'(t) + \lambda R(t) = \lambda N$$

- whose solution is:

$$R(t) = [R(t_0) - N]e^{-\lambda t} + N$$

- where R(t0) indicates the number of tweets at the initial instant.

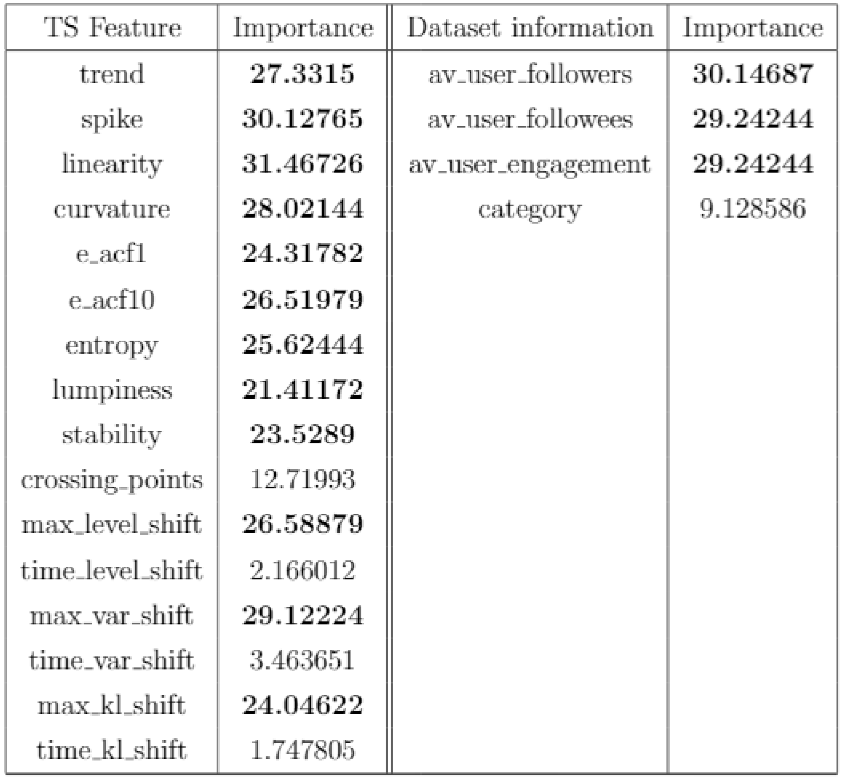
Carchiolo, V., Longheu, A., Malgeri, M., Mangioni, G., & Previti, M. (2018, October). Terrorism and War: Twitter Cascade Analysis. In *International Symposium on Intelligent and Distributed Computing* (pp. 309-318). Springer, Cham.

Data analysis: Fake news detection

- My approach consists on extracting features from the temporal evolution of each rumor and from the information about users involved in each rumor propagation.

Data analysis: Fake news detection

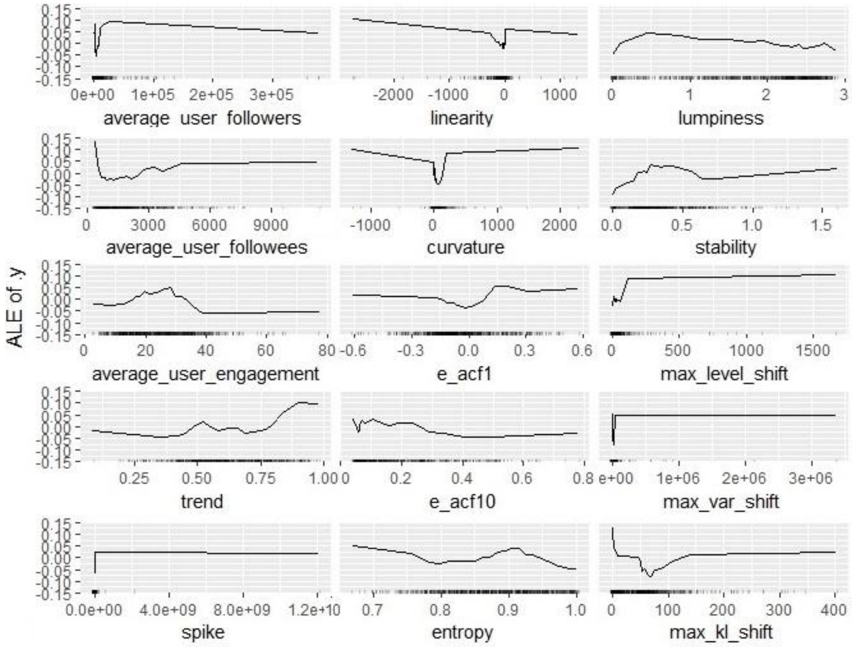
- The collection of time series features has been performed through the **tsfeatures tool**.
- I selected the ones that seemed the most promising, then, in order to have a confirmation about their relevance, after a pre-training of a random forest classifier with 500 trees and the rest of the parameters set to the default values given by the implementation of random forest used in the R package **mlr**, I computed the **Gini-importance** (i.e., the difference between a node's impurity and the weighted sum of the impurity measures of the two child nodes in tree) to remove the unsuitable features for our dataset.
- The random forest classification performed without less important features in 10-fold cross-validation have an accuracy of **84.61%**.



TS Feature	Importance	Dataset information	Importance
trend	**27.3315**	av_user_followers	**30.14687**
spike	**30.12765**	av_user_followees	**29.24244**
linearity	**31.46726**	av_user_engagement	**29.24244**
curvature	**28.02144**	category	9.128586
e_acf1	**24.31782**		
e_acf10	**26.51979**		
entropy	**25.62444**		
lumpiness	**21.41172**		
stability	**23.5289**		
crossing_points	12.71993		
max_level_shift	**26.58879**		
time_level_shift	2.166012		
max_var_shift	**29.12224**		
time_var_shift	3.463651		
max_kl_shift	**24.04622**		
time_kl_shift	1.747805		

Data analysis: Fake news detection

- In order to interpret the resulting model, I analyze an **Accumulated Local Effects (ALE)** plot, a common technique in the field of explainable machine learning.
- Looking at the effect of **user-based features**, false news spreaders usually have a large number of followers, a low number of followees and a low user engagements. This means that their accounts were created ad hoc to reach a particular category of users and only become active when a particular idea must be propagated, remaining inactive for the rest of the time.
- Looking at the ALE curves of **time series-based features,** for values close to 0 of *linearity* and *curvature*, the prediction is often true, while it is false for the rest of values of these features. That indicates that, unlike false rumors, true news tend to have a constant diffusion. Considering the above mentioned formula for *strength of trend*, for high variations of the ft component the rumor is classified as false, which hints that most of the false news have an higher variation of spreading over time with respect to true news. Finally, regarding the *max_level_shift* and *max_var_shift,* I see that, except for low value of these two features, the classifier gives false news prediction. True rumors tend to have fewer shifts in the evolution of number of tweets than fake news.

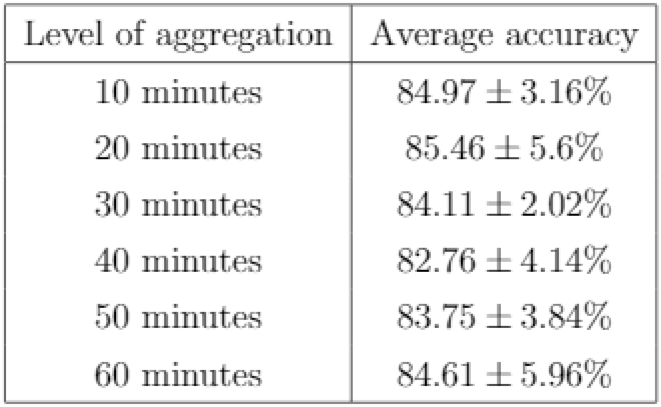


UAM
UNIVERSIDAD AUTONOMA
DE MADRID

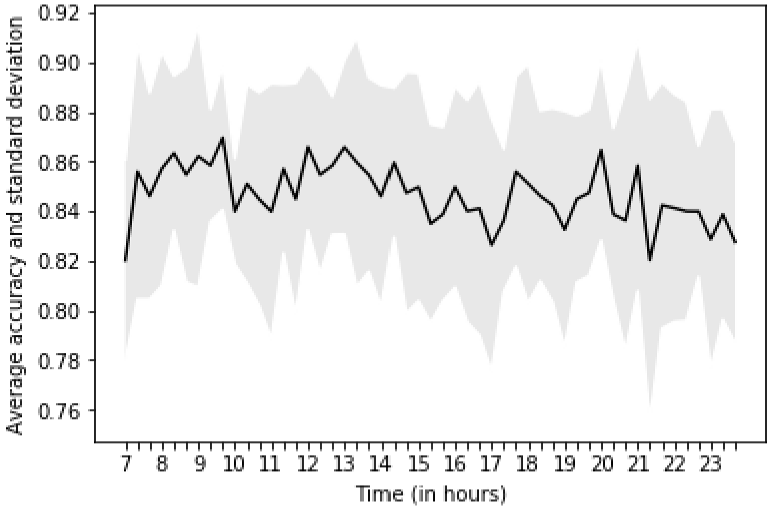
ai+da

Data analysis: Fake news detection

- I aggregated the tweets with a range of levels of aggregation, from 10 to 60 minutes, in 10-minutes increments and apply my methodology.
- Considering that the best result is found with 20-minutes samples, I carried out a study of the evolution of the accuracy with this level of aggregation, in order to understand if a smaller time can be considered without losing too much accuracy and discovered that there are not high variations in accuracy: the lowest values of accuracy in average is **82.02 +/- 4.06%** at 7 hours and 20 minutes, the highest **86.95 +/- 2.9%** at 10 hours.
- It happens because I also exploited information about users that embed the previous each other interactions of users, so, differently from similar time series approaches that need hours before to reach a stable values of accuracy, I have quite constant results in classification.



Level of aggregation	Average accuracy
10 minutes	$84.97 \pm 3.16\%$
20 minutes	$85.46 \pm 5.6\%$
30 minutes	$84.11 \pm 2.02\%$
40 minutes	$82.76 \pm 4.14\%$
50 minutes	$83.75 \pm 3.84\%$
60 minutes	$84.61 \pm 5.96\%$



Publication list

Credibility-based threshold model for not competitive and competitive news spreading on online social networks

Carchiolo, Longheu, Malgeri, Mangioni, Previti

(submitted to Complexis)

Fake news detection using time series and user features classification

Previti, Rodriguez, Camacho, Carchiolo, Malgeri

(submitted to EvoStar)

The impact of users trust on social networks news spreading

Carchiolo, Longheu, Malgeri, Mangioni, Previti

(submitted to SNAM journal)

Terrorism and War: Twitter cascade analysis

Carchiolo, Longheu, Malgeri, Mangioni, Previti

IDC 2018 - 12th International Symposium on Intelligent Distributed Computing

Bilbao, Spain, 15-17 October, 2018.

An Efficient Real-Time Monitoring to Manage Home-Based Oxygen Therapy

Carchiolo, Compagno, Malgeri, Trapani, Previti , Loria, Toja

WorldCist'18 - 6th World Conference on Information Systems and Technologies

Naples, Italy, 27 - 29 March 2018.

A trust based news spreading model

Carchiolo, Longheu, Malgeri, Mangioni, Previti

Complenet '18

Boston, Massachusetts, 5 - 8 May 2018.

Introducing credibility to model news spreading

Carchiolo, Longheu, Malgeri, Mangioni, Previti

Complex Networks 2017 - 6th International Conference on Complex Networks & Their Applications

Lyon, France, 29 November - 1 December 2017.

Post sharing-based credibility network for social network

Carchiolo, Longheu, Malgeri, Mangioni, Previti

IDC 2017 - 11th International Symposium on Intelligent Distributed Computing

Belgrade, Serbia, 11 - 13 October, 2017.

Monitoring students activities in CS courses

Carchiolo, Longheu, Previti, Fichera

15th RoEduNet Conference: Networking in Education and Research

Târgu-Mureș, Romania, September 7 - 9, 2016.

Tourism Websites Network: Crawling the Italian Webspace

Longheu, Mangioni, Previti

DATA NALYTICS 2016, The Fitfth International Conference on Data Analytics

Venice, Italy, October 9 - 13, 2016.

Schools

- *Mediterranean School on Complex Networks*

  Salina

  September 3-8, 2017
- *Complex Networks: Theory, Methods and Applications*

  Lake Como School of Advanced Studies

  May 14-18, 2018
- *International School on Informatics and Dynamics in Complex Networks*

  University of Catania

  October 15-19, 2018
- *International School on Data Science and IoT*

  University of Catania

  September 9-13, 2019

Thank you for your attention!