Characterizing the spread of a scientific rumour on Twitter

Leonardo Epifânio , **Pedro Lopes**
Instituto Superior Técnico
Av. Rovisco Pais 1, 1049-001 Lisbon
Network Science 19/20
Group 6
leonardo.epifanio@tecnico.ulisboa.pt, pedro.daniel.l@tecnico.ulisboa.pt

Abstract

A history that is certainly present in the annals of physics, is the announcement of the discovery of a Higgs boson-like particle in LHC, at CERN, as this is one of the ultimate scientific endeavours of the 21st century. In this paper, we present a study of the spread of this information and its flow through Twitter. In this first part of the project, we characterize the social network of users that were involved in the process of the rumour spread, before, during, and after the 4^{th} July 2012, which was the date of the official announcement.

1 Introduction

The Higgs boson is named after physicist Peter Higgs, who in 1964, along with five other scientists, proposed the Higgs mechanism to explain why particles have mass [Higgs, 1964]. The search for its existence has been among research priorities in the fields of physics for about 50 years, it being called the "God Particle" [Lederman and Teresi, 2006]. The year 2012 will be remembered as one of the most important years for physics, as on the 4^{th} July 2012, the Atlas and CMS collaborations confirmed the boson's existence [Aad *et al.*, 2012].

This breakthrough in science happened in an era of global online social media with no precedents. In social networks, a lot of people interacted, discussed and followed the news of the discovery in platforms such as Twitter. In this first part of the project, we explore the user network in Twitter that participated in the spread of this history. This includes references to *"Higgs", "LHC", "CERN"* in tweets and retweets, responses/mentions, and likes. The datasets [SNAP, 2015] were provided by [De Domenico *et al.*, 2013], and have been built after monitoring the spreading process on Twitter, before, during, and after the official announcement.

The objective of this work aims to explain the observed time-varying dynamics of user activities such tweets, replies, mention and retweets during the spreading of this scientific "rumour".

2 Dataset overview

The available dataset includes four directional networks that have been extracted from user activities in Twitter, as:

1. Retweeting (retweet network).
2. Replying to existing tweets (reply network).
3. Mentioning other users (mention network).
4. Friend/follower social relationships among users involved in the above activities (social network).
5. Information about activity on Twitter during the discovery of the Higgs boson.

We use the library *igraph* [1] which implements a collection of network algorithms that enable the analysis of complex graphs. Since the library is implemented in C, a language in which we are somewhat proficient, we attempted to use C. However, a lot of problems surfaced, from the installation, to the actual execution of the compiled executable. Due to this library not being managed by any OS package manager, we had to build it from source and install it in our system, and since it runs with dynamic linking, the linker, occasionally, had problems locating the library. Furthermore, programming in C proved to be more exhausting, and so we ended up shifting to Python, as it is simpler. Also, the Python's syntax helps in this kind of problems. We were apprehensive at first, switching to Python, mainly because of the speed. Albeit some lost of performance in IO and initialization, *igraph* runs the CPU intensive tasks in C, thus the speed penalty is negligible.

The analysis process for each of the first four graphs are composed of 3 steps:

1. Create graph from an edgelist file.
2. Calculate metrics over the graph.
3. Plot the results.

The *igraph* library, offers a large range of algorithms to calculate metrics. Albeit the known efficiency of the implemented algorithms, our dataset has a considerably large size and some metrics regarding centrality measures and diameter, took a huge amount of time in our laptops.

The four networks analysed encode the following:

- **Retweet network** corresponds to the graph of retweets, nodes are users and edges is the relation of retweet. If user A retweets from user B, then A has a directed edge to B. The network totalizes 256491 nodes and 328132 directed edges.

[1] https://igraph.org/

Metric	Number
Average clustering coefficient	0.1408
Triangles	83023401
Number of triplets	589654837
Fraction of closed triangles	0.002901

Table 1: Social network's clustering metrics

- **Reply Network** is the graph representing the interaction of reply to existing tweets about that include references to "Higgs", "LHC" and "CERN". It includes 38918 nodes, picturing the users, and 32523 directed edges, meaning that if user A is replying to a tweeter made by user B the, A has a directed edge to B.
- **Mention network** are the mentions made by users in their tweets. Users mentioned in retweets are also accounted as mentions. For instance, if user A retweets the tweet "Hello C D" sent by user B, then the following links are created: A to C and A to D, because users C and D can be notified that they have been mentioned in a retweet. It has 116408 nodes, symbolizing the users, and 150818 directed edges which encode the interaction.
- **Social network** of the authors of the tweets: the resulting graph is composed of 456626 nodes, corresponding to the users, and 14855842 directed edges, that represent the follower/followed relation between them (meaning that if node A has a directed edge to B then, A follows B).

The fifth data set has the information about all the activity on Twitter before, during and after the discovery of the Higgs' boson. It has the following structure:

```
userA userB timestamp interaction
```

Interactions can be **RT** (retweet), **MT** (mention) or **RE** (reply). Each link is directed. The user IDs are the same in all the data sets (1) to (5).

In this report, we present the anatomy of the spreading of this scientific rumour by analysing the provided data sets. We analyse the activity patterns of the individuals that tweeted about this discovery over the period taken into consideration. Then, we propose a model for the information spreading over the Twitter network, considering the influence of the structure of the network on the process.

The developed code for the analysis and graphic plotting will be available in the delivered zip.

3 Results

3.1 Social Network

Degree distribution and Assortativity

We started by calculating the degree distribution. As it is a directed graph, we divided the analysis in 3 categories: in-degree (how many followers a certain user has), out-degree (how many users a certain user follows) and total-degree (the summation of both). After calculating and plotting the results on *matplotlib* [2], we got similar results as in [De Domenico *et al.*, 2013] for the in-degree, but slightly different fitting for the out-degree and total-degree. The results are shown in figure 1.

[2]https://matplotlib.org/

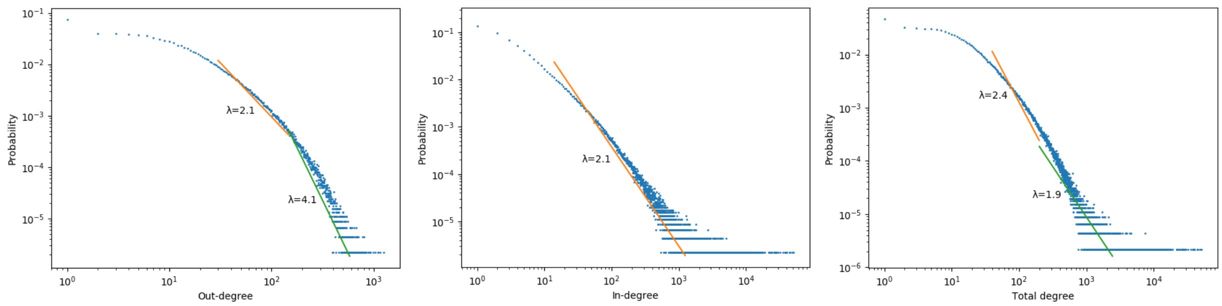
The underlying topology is not trivial and it shows a strange behaviour, especially for the out-degree distribution. All of the three showcase power-law scaling, however, for the out-degree distribution in figure 1a, it scales in two different regimes $P_{kout} \propto k_{out}^{-2.08}$ and $P_{kout} \propto k_{out}^{-4.12}$ with crossover $k_{out} \approx 150$, way beyond the extreme spectrum of a scale free network lambda, it presented an average degree of 32.53 and variance 2414.38. With this result it indicates that very few users follow more than 150 users, and after this point it starts to behave like a random network. This represents a challenge, to understand how users decide to follow a certain number of people in Twitter. Oppositely, the in-degree shown in figure 1b, shows a more understandable behaviour scaling in a single regime $P_{kin} \propto k_{in}^{-2.1}$, resulting in an average degree of 32.53 and variance 1.28×10^5. The number of followers that a user has falls in the normal interval regime of a scale-free network. For the total-degree, *igraph* both out-degrees and in-degrees are summed. The original paper replaced symmetric edges by only one, undirected, edge, thus getting different power-law fits. We'd like also to mention that the method for the powerlaw fit *igraph* implements follows [Clauset *et al.*, 2009].

One standard metric to find correlation in the network is by calculating assortativity. Nodes in a network with a large number of links may be connected to nodes with many connections (assortative, with a positive metric) or to nodes with low connections (disassortative, with a negative metric). In case of social networks, usually, the first prevails. Anyhow, we calculated the metric and the result was -0.14, a really strange value, indicating a disassortative correlation. One possible explanation for this result, is the fact that this graph is a sub-graph of, exclusively, users that mentioned one of the keywords in one of the activities explained in section 2. Thus, it might exhibit more dissasortative links than the original Twitter's full network where no topic restriction exists. This might suggest that, at least for this specific topic, information exchange between high-degree nodes(information hubs) and low-degree nodes (information consumers) prevails with significance.

Clustering

The clustering metrics allow us to obtain results over the number of present triangles, clustering coefficient and transitivity, enabling the assessment of the existence of highly connected groups in the network, and to draw conclusions about the resilience and robustness of the network.

The table 1 summarizes the obtained results, showing a total of 83023401 triangles and the average clustering coefficient 0.1408, meaning that there are a total of 589654837 triplets (open and closed). The clustering coefficient measures the degree to which nodes tend to cluster themselves. In the case of this graph, it presents quite a low clustering coefficient, thus it has no clustering. One explanation is, again, by this graph representing a population restricted by topic, and it gets a strange topology.



(a) Out-degree distribution. (b) In-degree distribution. (c) Total-degree distribution.

Figure 1: Social network's degree distribution of in-degree, out-degree and total-degree of the nodes that tweeted about Higgs boson

Metric	Number	Percentage
Nodes	456626	-
Edges	14855842	-
Nodes in largest WCC	456290	99.9%
Edges in largest WCC	14855466	100.0%
Nodes in largest SCC	360210	78.9%
Edges in largest SCC	14102605	94.9%

Table 2: Social Network's Strong and Weak Components

Strong and Weak Components

Regarding strongly connected components (SCC), these are described as sub-graphs where there is a path from every node to every node. In this case, there is a follower path from one user to any other user. We present the largest SCC that has a total of 360210 nodes (78.9% of total nodes) with 14102605 edges (94.9% of total edges). This is relevant, because it explains how the information flows in social platforms like Twitter, these metrics allow us to hypothesise that relatively few people are needed to spread the rumour. The largest SCC represents 78.9% of the network.

The weakly connected component (WCC) is one which all components are connected by some path, ignoring direction. In our results, the largest has 456290 nodes (99.9% of total nodes) and 14855466 edges (100.0%), which is almost the entire network. These results imply that almost all the network is connected with a relationship of follower/followed, strengthening the hypothesis above. Table 2 summarizes the results.

Shortest Path Measures

The algorithms which *igraph* implements for the calculation of these metrics have an overall complexity of $O(\mid V \mid\mid E \mid)$. The number of nodes is relatively small, but the number of edges ascends to almost 15 millions, making the calculation of the shortest path related measures a time consuming tasks that took on average 27 hours to finish. In the code, there is a warning when these metrics are about to get calculated. We ran the various algorithms in parallel to speed up the process as each one is independent. Table 3 summarizes our findings.

3.2 Activities Networks

We applied the same study for the other graphs that encode the activities and the way they connect the users. The resulting metrics are presented in Tables 4 to 6 and the degree distributions and its fittings are shown in Figure 3.

Metric	Number
Diameter	9
Average Path length	3.7

Table 3: Social Network's Shortest Path Measures

We can not extract relevant conclusions from those metrics: according to the degree most of the graphs behave like a random network where $P_k \propto k^{-\lambda}$ has $\lambda > 3$ with large variance. The exception here is the replies network where the opposite happens, we have the value $\lambda < 2$ and enter in the realm full of anomalies. This might be explained by the low activity the reply has in comparison with the other activities that are more dominant in Twitter.

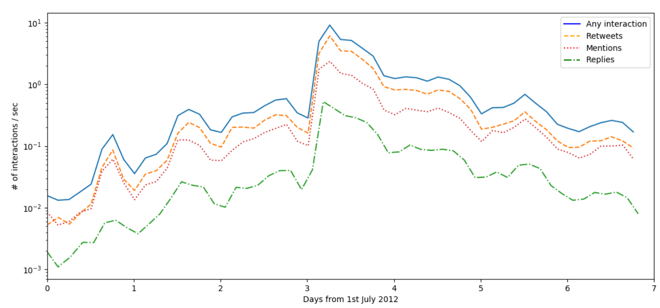


Figure 2: Number of activities per second as a function of time during the period of collected data.

A new approach followed was the study of evolution of the rate of activities made by the users, pictured in fig. 2. A rate shows a rapidly increasing trend up to the day of the announcement by the CERN team, followed by a slowly decrease. It is worth noting that, when all activities are considered, the rate of activities increases from approximately 76 activities/hour at the beginning of Period I up to 36000 activities/hour on its peak. The rumours anticipating the presentation of results at Tevatron caused the initial spreading of tweets about the Higgs boson. This was further sustained by the subsequent comments to these initial postings and the rumours about the results to be presented by the scientists belonging to the ATLAS and CMS experiments. During a few hours after the announcement of the discovery, the rate increased by more than one order of magnitude, while it slowly decreased in the following days.

Metric	Retweet	Reply	Mention
Diameter	9	26	10
Average Path length	2.9	11.0	2.9

Table 4: Activities Networks' Shortest Path Measures

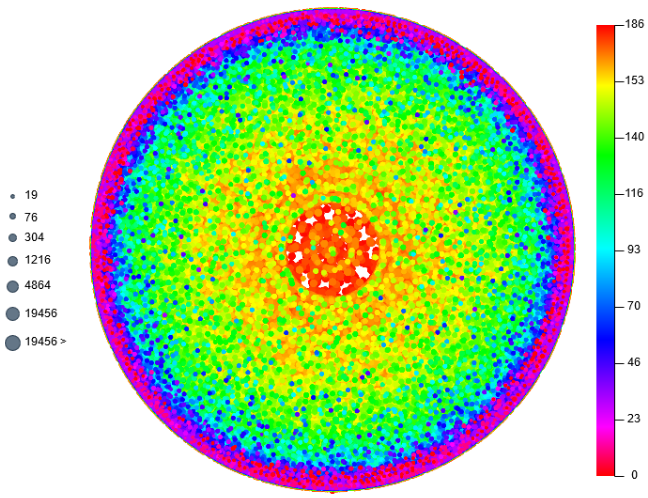


Figure 4: Social network visualization based on k-core decomposition and components analysis. The size of each vertex is proportional to its degree and the color encodes the k-coreness

4 Rumour Spreading

In this section, we investigate the dynamics of information spreading in the social network of users who tweeted about the Higgs boson. To investigate these dynamics we make use of epidemic models, originally used to study infection spreading, since there can be found some similarities between the spreading of information and of diseases. Disease epidemics usually depend on physical contact between individuals, influenced by the multitude of differences in the biological characteristics of the individuals. However, information spreading does not need physical contact to be spread. Communication infrastructures such as television, newspapers, radio and the Internet are used instead to spread this information, which is very volatile, since individuals do not incubate it, spreading it based on their decisions.

There are many models for disease spreading, the simplest being the SI model where the individual can be in two possible states: Susceptible (S) and Infected (I). The individuals are then able move between these states, depending on probabilities regarding infection and recovery.

In this work, we apply the same contagion methodology but rename the states to Active user and Non-active user and indicate them with $A(t)$ and $D(t)$, respectively, with $A(t) + D(t) = N$, where N represents the total number of users in the considered network. Rumour spreading is then represented by the user-to-user interaction processes, described in the activity datasets. When a user performs an activity (RT, MT, or RE), he is considered active, and therefore, a rumour spreader.

We can observe the social network shown in Figure 4, where a visualization based in the k-coreness decomposition and component analysis is presented. Vertices with the highest k-coreness act as the most influential spreader of information in the network. In fact, it has been recently shown that in some plausible circumstances the best spreaders are not the most highly connected or the most central people but those with higher k-coreness, and there is evidence of a positive correlation between k-coreness and the size of cascade messages suggesting that users at the core of the network are more likely to be the seeds of global chains of information diffusion.

In Figure 4, there is vertices with high-degree in any k-shell, meaning a very low correlation between degree and the shell-index, indicates that hubs are likely to be found also in external shells, a behaviour typical of network without an apparent global hierarchical structure, similar to the World Wide Web.

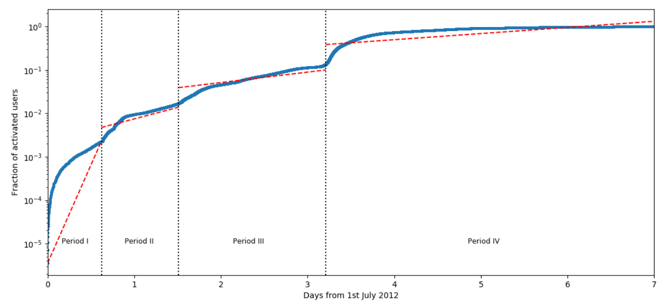


Figure 5: Fraction of users that are active atleast once with respect to the total number of users taken into consideration. Lines indicate the fitting results obtained separately for each temporal range by adopting the model SI (eq. (1)).

Modeling the dynamics of user activation without de-activation

To analyze how users become rumour spreaders over time, we use the activity data set (referenced as the 5^{th} in section 2), and first study a model which doesn't consider user deactivation. In this model, a user is considered activated when it has performed its first activity in the time period analyzed, and, since there is no de-activation, the user stays activated until the end of the period.

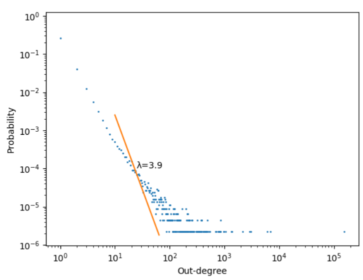
Hence, the fraction of active users, can be represented as a monotonic increasing function of time, $a^*(t)$. As can be seen in Figures 5 and 6 we divide time into four periods, each related with a phase concerning the spreading of the scientific rumour. Starting on the July 1^{st} 2012 at 02:52 UTC, the analysis starts on **Period I**, which represents the first rumours. **Period II** represents the time when the results were announced at Tevatron, followed by **Period III**, which features a more gradual incline in the number of active users, until **Period IV**, which starts about when results were announced by the LHC, and continues up to July 8^{th}.

Metric	Retweet	Reply	Mention
Nodes	456623	456612	456620
Edges	492198	32523	226227
Nodes in largest WCC (%)	244811 (53.6%)	12839 (2.8%)	108672 (23.8%)
Edges in largest WCC (%)	486030 (98.7%)	14944 (45.9%)	221457 (97.7%)
Nodes in largest SCC (%)	39136 (8.6%)	322 (0.1%)	21936 (4.8%)
Edges in largest SCC (%)	249515 (50.7%)	708 (2.2%)	118802 (52.5%)

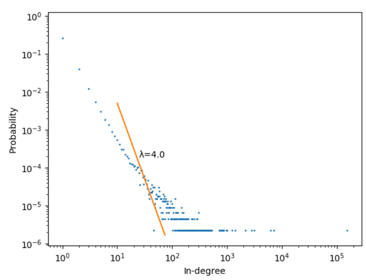
Table 5: Activities Networks' Strong and Weak Components

Metric	Retweet	Reply	Mention
Average clustering coefficient	4.9×10^{-2}	3.7×10^{-4}	2.8×10^{-2}
Triangles	21172	244	23068
Number of triplets	432082	659459	809404
Fraction of closed triangles	1.09×10^{-4}	1.56×10^{-4}	2.42×10^{-4}

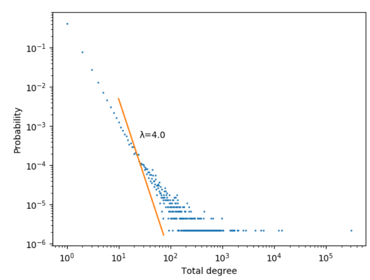
Table 6: Activities networks' clustering metrics



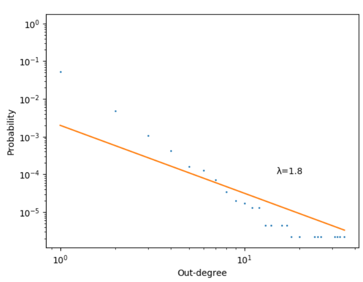
(a) Retweet Out-degree distribution.



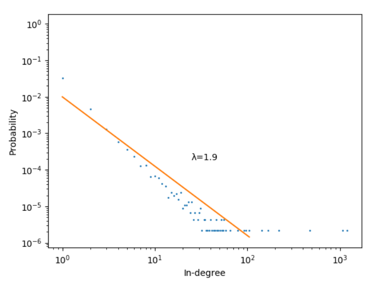
(b) Retweet In-degree distribution.



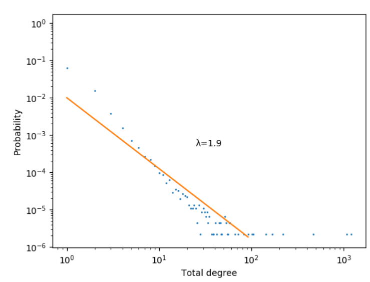
(c) Retweet Total-degree distribution.



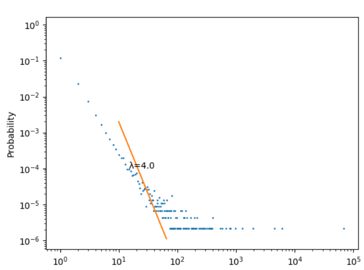
(d) Reply Out-degree distribution.



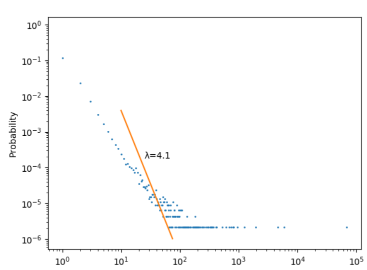
(e) Reply In-degree distribution.



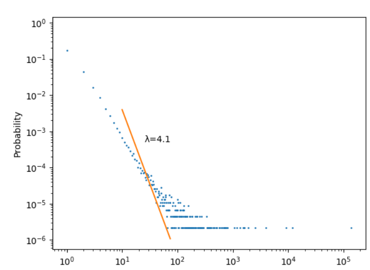
(f) Reply Total-degree distribution.



(g) Mention Out-degree distribution.



(h) Mention In-degree distribution.



(i) Mention Total-degree distribution.

Figure 3: Activities networks' degree distribution of in-degree, out-degree and total-degree of the nodes that made an interaction about Higgs boson

We first tried to model the evolution of the fraction of active users using the simple SI model, with the equation:

$$a^*(t) = a_0^* \times e^{t/(\lambda \times k)^{-1}} \tag{1}$$

in which λ is seen as the activation rate, and k is the average degree. The results are displayed in Figure 5, and as can be seen, in spite of any effort to adjust the activation rate, this model leads to a very poor fit of the data, even when trying to apply different fits to the several periods.

To solve this issue, we then used another new model, developed in [De Domenico *et al.*, 2013], built upon an observation of the dataset, which states that, generally, once a user has had activity, he will not do so significantly in the near future, according to the bursty behaviour that can be seen in the results. And so a simplifying assumption is made, that a user will not be active again after his first interaction.

The number of newly active users A^* at time t is then proportional to the before inactive users:

$$A^*(t + \Delta t) = A^*(t) + \lambda^*[N - A^*(t)]\Delta t, \tag{2}$$

where λ^* is the activation rate, and N is the total number of users. In the limit of a small Δt, the following is obtained:

$$\frac{da^*(t)}{dt} = \lambda^*[1 - a^*(t)], \tag{3}$$

and so, the evolution function, $a^*(t)$, is the solution of Equation (3), and is calculated by:

$$a^*(t) = 1 - [1 - a^*(t_k)]e^{-\lambda^*(t - t_k)}, \tag{4}$$

in which t_k is the starting time of period k.

By tweaking the activation rate parameter in each period, we managed to obtain a fit that finely models the data, as is shown in Figure 6, where we obtained, for Period I, an activation rate of 0.192 users/min, for Period II, 1.008 users/min, followed by a rate of 3.852 users/min in Period III, which is 20 times the activation rate for Period I, which makes sense, since this is after the announcement at Tevatron. In Period IV, with the announcement by the LHC, there is a burst in tweets related with the matter, and the period is characterized by an activation rate of 83.058 users/min.

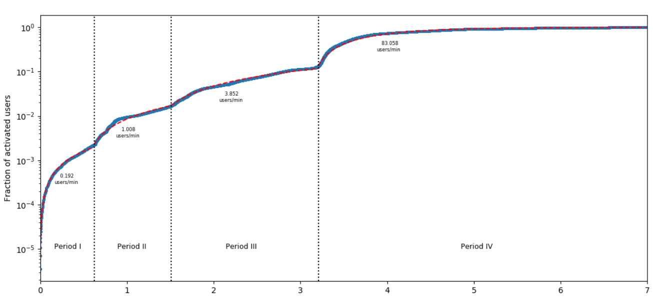


Figure 6: Fraction of users that are active atleast once with respect to the total number of users taken into consideration. Lines indicate the fitting results obtained separately for each temporal range by applying eq. (2).

These results differ from the ones in [De Domenico *et al.*, 2013], where the activation rates are consistently larger (in average, 6 times larger, with little variance), despite the resulting fit being very similar. This might be explained by the fact that they claim to have datasets relating to tweets, over time, that use certain keywords associated with the Higgs boson context, whilst we only have access to the activity that concerns interactions between users. This way, they may have more users involved in the rumour spreading, and if the assumption is right, it is interesting to notice how, even with this difference, the fraction of active users in both cases evolves in a very similar way.

Modeling the dynamics of user activation with de-activation

In this section, the dynamics are analyzed using a refined model, based on SIS, where an active user may become non-active again, which occurs when, in a given time window, Δt, the user hasn't had any activity. This de-activation factor, in our context, can also account for the limited visibility of Higgs boson related tweets, which may be replaced by newer tweets in the timeline. To denote the probability, over time, of an active user changing to the non-active state, we use $\beta(t)$.

If a non-active user is connected to at least one active peer, there is a larger probability of this user becoming active. Using j_A to represent the number of active users in the neighbourhood of a non-active user, and $\lambda(t)$ the activation probability per unit of time t, the probability of a user becoming active is given by $p_\lambda(t; j_A) = 1 - [1 - \lambda(t)]^{j_A}$.

Since it has been shown, in previous work, that correlations at neighbourhood level do not affect the spreading dynamics, these are disregarded, and the probability, over time, of a non-active user, with in-degree k^in, being connected to j_A active users is represented by:

$$\tilde{p}(t; j_A, k^{in}) = \frac{\binom{A(t)}{j_A}\binom{N-A(t)-1}{k^{in}-j_A}}{\binom{N-1}{k^{in}}}, \tag{5}$$

which is essentially the combinations of the number of activated users with j_A, multiplied by the combinations of the users in the neighbourhood which are inactive, from the total number of combinations from the remaining $N - 1$ users with k^{in}.

Then, the probability of a user with k^{in} becoming active thanks to, at least, one of his neighbours, is given by:

$$P_{\lambda,k^{in}}(D \longrightarrow A) = \sum_{j_A=1}^{k^{in}} \tilde{p}(t, j_A, k^{in}) p_\lambda(t; j_A), \tag{6}$$

Therefore, the total probability of non-active nodes becoming active is

$$\Theta_\lambda(t) = \sum_{k^{in}} \mathrm{P}(k^{in}) P_{\lambda,k^{in}}(D \longrightarrow A), \tag{7}$$

in which $\mathrm{P}(k^{in})$ is the probability density of the in-going degree. The number of active users, in the time interval Δt, is then modelled by:

$$A(t+\Delta t) = A(t) + \left[-\beta(t)A(t) + (N - A(t))\Theta_{\lambda(t)}(t)\right]\Delta t. \tag{8}$$

Using as the time interval, one single time unit, $\Delta t = 1$, the model becomes:

$$A(t+1) = \left(1 - \tilde{\beta}\right) A(t) + (N - A(t))\Theta_{\tilde{\lambda}(t)}(t). \quad (9)$$

Considering the density of active users as $\rho(t) = A(t)/N$, the evolution of this density is given by:

$$\rho(t+1) = (1 - \beta(t))\rho(t) + (1 - \rho(t))\Theta_{\lambda(t)}(t). \quad (10)$$

Because of the complexity of $\Theta_{\lambda(t)}(t)$, the model was modified by introducing a variable activation rate λ, that varies in time, accounting for the decreasing interest, according to:

$$\lambda(t+1) = (1 - \xi)\lambda(t), \qquad 0 < \xi < 1, \quad (11)$$

We interpret ξ as the inverse of characteristic scale τ regulating the decay dynamics. Which, in other words, is the natural time scale over which interest fades.

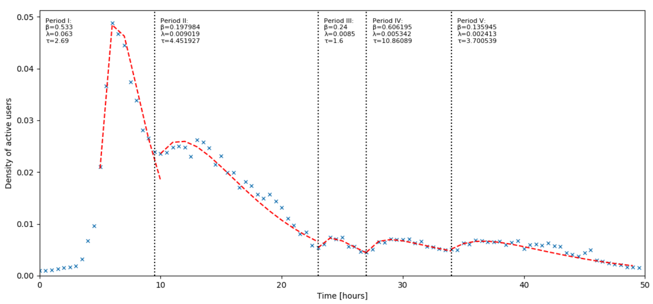


Figure 7: Observed evolution of the density of active users versus time (points) in Period IV, i.e., during and after the main event, from 00:00 AM, 4^{th} July. Curves indicate the predictions obtrained from the model defined by eq. (10) and eq. (11), where the values of the corresponding parameters are reported in the figure for different sub-periods. The reported λ refers to the initial value of the activation rate.

The model was applied only to Period IV, where we identify five sub-periods, each one characterized by an increase of active users followed by a decreasing one. The curves presented in fig. 7 correspond to the model defined by eq. (10) and eq. (11), by solving the parameters λ, β and τ to minimize χ^2, in order to fit the obtained data for each sub-period. The first rapid increase in the first sub-period is followed by a fast decrease, with time scale $\tau \approx 2.69$ hours, initial activation rate $\lambda_0 = 0.063$ and $\beta = 0.533$. The fact that the density remains steady in the first hours of the period is due to the fact that the period starts before the announcement, which accounts for the observed variance. In the next periods, the number of active users decreases somewhat steadily, until stagnation.

5 Conclusion

In the first part of the study we placed more focus on the topology of the network of users that interacted in the news spreading about the discovery of the Higgs boson. Due the considerable size of the network, we focused mainly on the shortest path related metrics. The algorithm that *igraph* implements to obtain those metrics has an overall complexity of $O(|V||E|)$. Despite the long time, we were able to get the metrics. We would also like to mention, that the load of the graph in memory and the execution of the algorithms over it didn't take too much memory, just a few hundred megabytes.

After the characterization of the social network graph, we monitored the activities of the users before, during and after the main event which was the official announcement of the discovery. The rates of activities per second showed a rapidly increase trend up to the day of the announcement, after which slowly decreases.

The SI and SIS models were not good enough to the obtain the data fits. A new model was necessary for this particular information spreading with variable activation rate. We analyse the model in two dynamics: with and without activation. The model assumes memoryless individuals, where the activation is obtained by social reinforcement at the neighbourhood level. Activating the transition from non-active to active from a user has a larger probability if most of his/her friends are tweeting the rumour repeatedly in time. The provided model is really close to the obtained data, and we were even able to enhance the results obtained in [De Domenico *et al.*, 2013].

Accounting the pratical analysis, overall, the *igraph* library was pretty easy to use, and its documentation very complete, making our developing experience normal without big bumps. Even in the creation of the graph, we didn't had the necessity to pre-process, since the original data set was in a known format making it faster.

All the graphics in this report can be obtained by executing the delivered code.

References

[Aad *et al.*, 2012] Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, O Abdinov, R Aben, B Abi, M Abolins, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.

[Clauset *et al.*, 2009] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[De Domenico *et al.*, 2013] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3:2980, 2013.

[Higgs, 1964] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, 1964. [,160(1964)].

[Lederman and Teresi, 2006] Leon M Lederman and Dick Teresi. *The God particle: if the universe is the answer, what is the question?* Houghton Mifflin Harcourt, 2006.

[SNAP, 2015] Stanford University SNAP. Higgs Twitter Dataset. https://snap.stanford.edu/data/higgs-twitter.html, 05 2015. Accessed: 2019-11-27.