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 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 first part of the project is to get ourselves acquainted with the tools to analyze graphs and extract their properties.

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 explore just the item 4, which is a directed, unweighted, graph. 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 task intensive in C, thus the speed penalty is negligible.

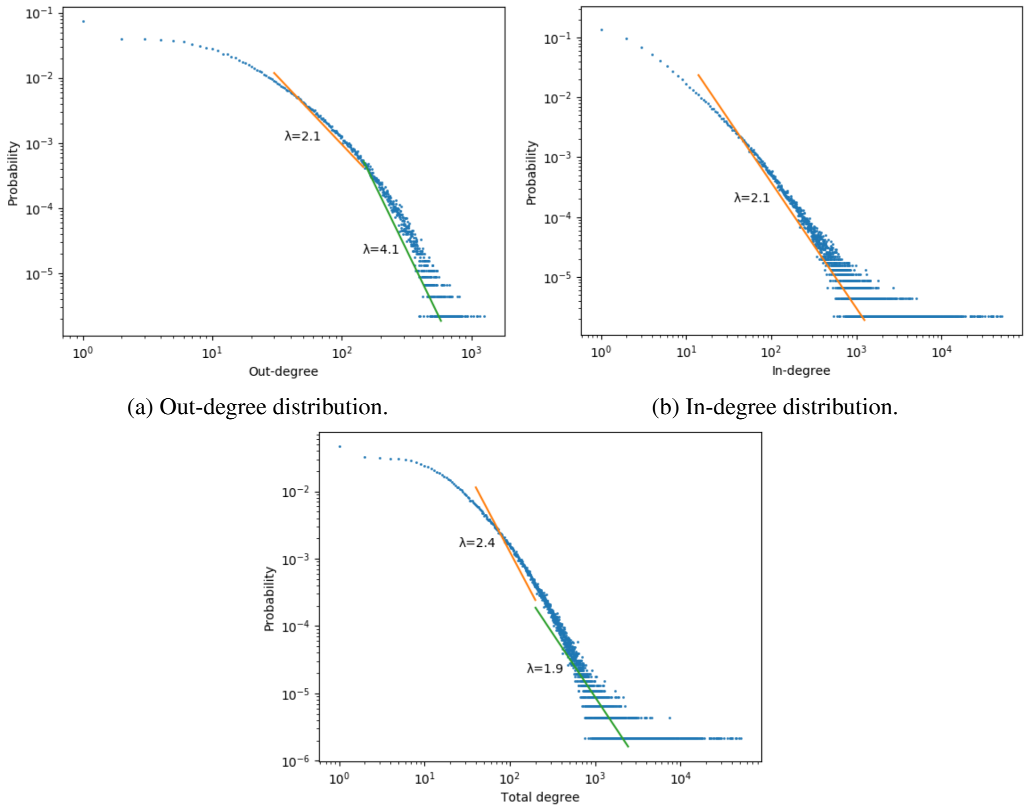
The analysis process is composed of 3 steps:

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 network we analysed is the 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 developed code will be available in the delivered zip.

[1] https://igraph.org/



(a) Out-degree distribution.

(b) In-degree distribution.

(c) Total-degree distribution.

Figure 1: Degree distribution of in-degree, out-degree and total-degree of the nodes that tweeted about Higgs boson

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

Table 1: Clustering metrics

3 Results

3.1 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.

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

[2] https://matplotlib.org/

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: Strong and Weak Components

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.

3.2 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.

3.3 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.

3.4 Shortest Path Measures

The algorithms which *igraph* implements for the calculation of these metrics have an overall complexity of $O(\mid V \mid\mid E \mid)$.

Metric	Number
Diameter	9
Average Path length	3.7

Table 3: Shortest Path Meaasures

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.

4 Conclusion

In this first part of the project 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. At the beginning we ran into a lot of problems using *igraph*, mainly because of our stubbornness in using C. When we changed to Python, the workflow got smoother.

Due the considerable size of the network, there were some metrics that we were eager to get, but took a long time, mainly in the shortest path related metrics. The algorithm that *igraph* implements to obtain those metrics has an overall complexity of $O(\mid V \mid\mid E \mid)$. 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.

The *igraph* library was pretty easy to use it, and its documentation is 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, the original dataset was in a known format making it faster.

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.