Università di Pisa

Social Network Analysis
A.A. 2017/2018

Cambridge Analytica and Facebook: The Scandal and the Fallout on Twitter

Gianmarco Ricciarelli 555396
Stefano Carpita 304902

Data drives all we do.

Cambridge Analytica main slogan.

Rules don't matter for them.
For them, this is a war, and it's all fair.

Christopher Wylie,
former datascientist at Cambridge Analytica, about its leaders.

Contents

1 | The case story

On Saturday 17^{th} of March 2018, The New York Times and The Guardian / The Observer broke reports on how the consulting firm Cambridge Analytica harvested private information from the Facebook profiles of more than 50 million users without their permission, making it one of the largest data leaks in the social network's history [1], [2].

Cambridge Analytica described itself as a company providing consumer research, targeted advertising and other data-related services to both political and corporate clients. The whistleblower Christopher Wylie, datascientist and former director of research at Cambridge Analytica revealed to the Observer how Cambridge Analytica used personal information taken without authorisation in early 2014 to build a system that could profile individual US voters, in order to target them with personalised political advertisements. Christopher Wylie, who worked with a Cambridge University academic to obtain the data, told the Observer: "We exploited Facebook to harvest millions of people's profiles. And built models to exploit what we knew about them and target their inner demons. That was the basis the entire company was built on."

We report here a timeline of the events of the first days of the scandal, from different sources [3]:

- March 17: The Observer and The New York Times publish joint reports on data harvesting by Cambridge Analytica. UK Information Commissioner Elizabeth Denham issues statement that they are "investigating circumstances in which Facebook data may have been illegally acquired and used." Politicians in US and UK call for investigation.
- March 19: Channel 4 News publishes part 1 of their undercover investigation into Cambridge Analytica. Facebook sends investigators to Cambridge Analytica's offices. UK Information Commissioner orders them to stand down. [4]
- March 20: Channel 4 News publishes part 2 of their undercover investigation into Cambridge Analytica, where they boast about getting Donald Trump elected. British MP Damian Collins calls on Facebook to present oral evidence on Cambridge Analytica. Facebook agrees to send former operations manager Sandy Parakilas. Facebook holds internal Q&A with attorney Paul Grewal to discuss the crisis, but CEO Mark Zuckerberg and COO Sheryl Sandberg do not attend. Cambridge Analytica suspends CEO Alexander Nix. Facebook demands to inspect Christopher Wylie's phone. FTC opens investigation into Facebook.
- March 21, Brian Acton, co-founder of the messaging app WhatsApp, called on his Twitter followers to #deletefacebook. Facebook purchased WhatsApp in 2014.

  @brianacton: It is time. #deletefacebook

  In a Facebook post published several days after the initial reports, Zuckerberg responded to the continued fallout over the data scandal. He said:
  "We have a responsibility to protect your data, and if we can't then we don't deserve to serve you. I've been working to understand exactly what happened and how to make sure this doesn't happen again."
- March 23, Elon Musk joined the #DeleteFacebook movement, taking down official pages for two of his companies, Tesla and SpaceX, announcing his decision on Twitter. Both the SpaceX and Tesla accounts vanished within hours of his tweet.
- March 24, The Ars Technica website reported that Facebook "surreptitiously" collected call and SMS data for years from Android users, including names, phone numbers, and the length of calls [5].

The described case arose a wide interest in privacy protection and regulation, offering an unique chance of public debate. On the 25^{th} of May 2018 the European General Data Protection Regulation (GDPR), regarding the privacy protection of all the individuals within the European Union, became effective after its approval two years before [6]. This event constitutes an important step for the safeguard of digital rights, although not all the issues have been treated and proposals of stricter treatments about privacy issues are on the table [7]. But, beside laws and rules, the most important protection against abuses or misuses it is the ethical and responsible behaviour of data scientists, stakeholders and all the people involved [8].

2 | Building the network

In this report we present an analysis of the spreading of the Cambridge Analytica-Facebook scandal on Twitter. We have considered a network composed by the authors of tweets about the case, during the first period of the scandal outbreak. The data have been collected via the Twitter API and we built the network using the following consecutive steps:

1. Crawling of all the available tweets over a period of more than 15 days, since the 17^{th} of March, containing at least one of the most popular hashtags regading the case:
   {#cambridgeanalytica, #facebookgate, #deletefacebook, #zuckerberg }
2. Cleaning of the crawled tweets, by selecting and storing in a MongoDB database only the users informations about the authors of tweets, excluding retweets and mentions.
3. Selection of the case outbreak time period by observing the time history in Fig. 2.1. The selected time period consists of 8 days, from the 17^{th} to the 24^{th} of March included (considering the Italian timezone).
4. Crawling of the following list for each of the selected authors, extracting the followee/follower relationships.

In Fig. 2.1 is represented the rate of new authors per hour. We can observe a tipical daily periodicity, with peaks during the Italian afternoon, corresponding to the USA waking up. The first relevant rate increase is observed on the Monday 19^{th}, following the weekend of news publishing. A stationary evolution is observed after the 24^{th} of March, a week after the scandal outburst.

The collecting of the following lists of all the authors has been necessary in order to overcome the huge time complexity of a direct crawling of the followee/follower relationships, considering also the Twitter API rates limits. A negligible fraction of the following lists was not available, because of the Twitter privacy settings of the users.

In summary, we built a directed unweighted network, consisting of:

- **Nodes**: 65729 twitter users, authors of tweets containing at least one of the hashtags specified above.
- **Edges**: 2501757 followee/follower relationships between the selected authors.

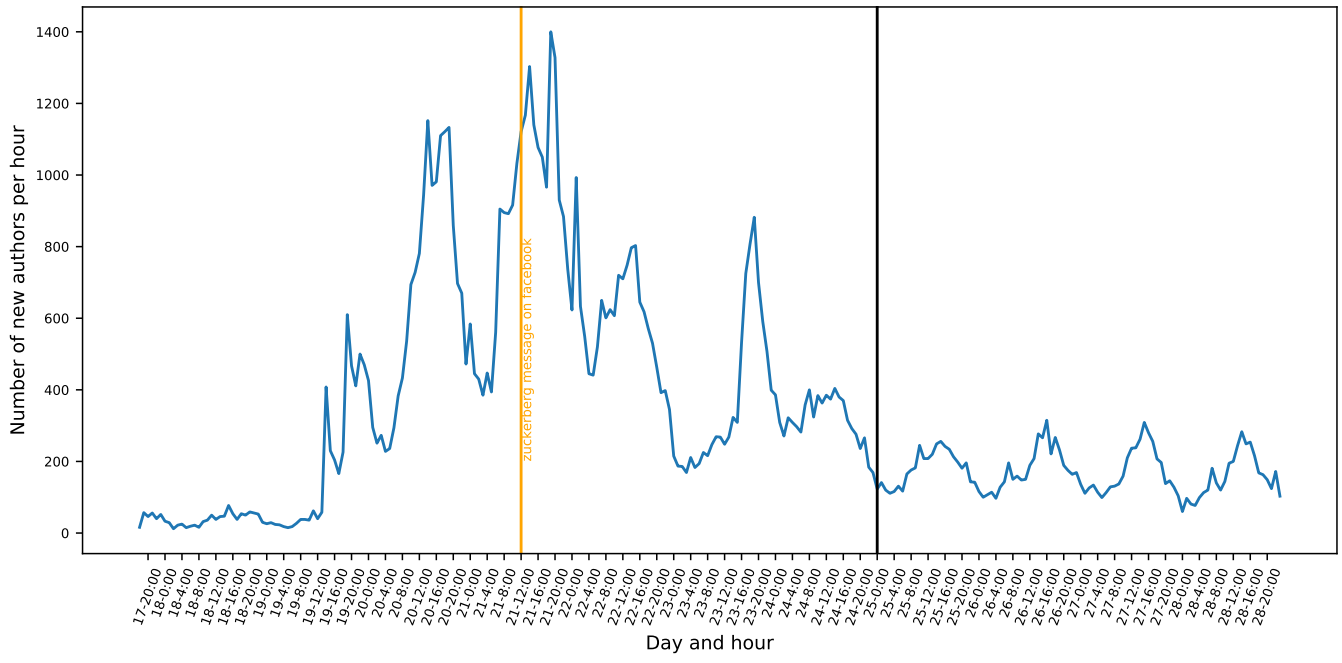


Figure 2.1: Time history of the rate of new authors per hour, during the first 12 days from the first publishing of the scandal.

3 | Network analysis

3.1 Network properties summary

The properties of the networks analyzed are summarized in Tab. **??**. The g network represents the original directed graph, while g_{und} is the graph obtained by not considering the directivity of the links. We observe that the number of links of g_{und} is lower respect to the one of g: the difference of 605879 represents the number of reciprocal links, corresponding to the number of user pairs following each other, a 24% fraction of the overall number of authors.

The network has been compared throughout the report with two synthetic network, an Erdos-Renyi random network g_{ER} and a Barabasi-Albert network g_{BA}. The Erdos-Renyi random network, with directed connections, has been generated with a value of the "linking probability" p computed using the double of the average degree of the original network:

$$p_{ER} \approx \frac{2\langle k \rangle}{N} = \frac{76}{65729} \approx 0.001 \tag{3.1}$$

An undirected network built with the Barabasi-Albert model has an average degree equal to the double of the links formed by each new node: $\langle k_{BA} \rangle = 2m$ [9]. Because our network is directed we generated a Barabasi-Albert random network with a value of m equal to half of the double of the average degree of the original network, or simply equal to the average degree of our directed graph:

$$m = \frac{2\,\langle k \rangle}{2} = \langle k \rangle = 38 \tag{3.2}$$

The network g_{Ita} it is the subgraph of g including only the Italian users, who have been identified by using the language metadata available with the crawled tweets.

	g	g_und	g_er	g_ba
L	2501757	1895878	4318406	4989628
N	65729	65729	65729	65729
density	0.00058	0.00088	0.001	0.00231
gamma	2.6	None	None	2.9
gamma_in	2.4	None	None	None
gamma_out	2.9	None	None	None
gamma_tot	2.6	None	None	None
k_avg	38	57	65	151
k_in_max	19064	None	109	None
k_in_min	0	None	36	None
k_max	19073	19065	183	3640
k_min	1	1	84	76
k_out_max	4130	None	103	None
k_out_min	0	None	35	None

Table 3.1: Label

The network is composed by a large strongly connected component including about the 75% of the original graph, as depicted in Fig. 3.1. The remaining nodes represents users who follows only few authors of the graph. Without considering the directivity of the links we have a weakly component including almost the whole network. The generated random graphs are completely connected.

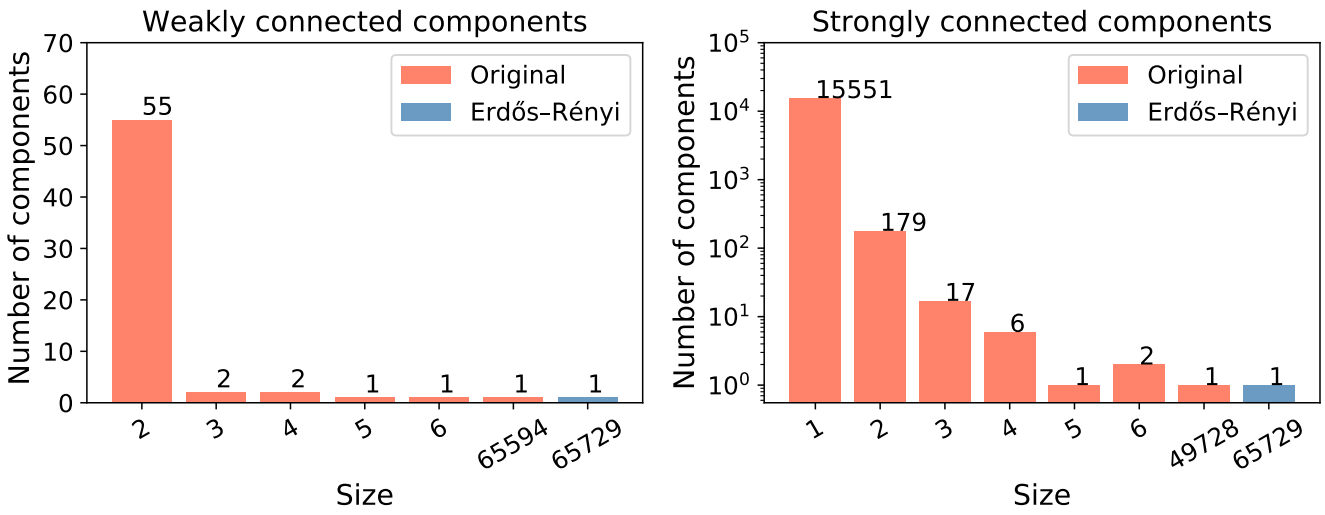


Figure 3.1: Weakly and strongly connected components.

3.2 Degree distributions

The analysis of the degree distribution of a network permits to identify the characteristic power law trend, immediately recognizable by using a log-log scale plot. Observing Fig. 3.2 it is evident the difference between the degree distributions of the original graph respect to the random generated Erdos-Renyi graph. The log scale used requires to use a logarithimc binning of the degrees values: in Figs. 3.3, 3.4, 3.5, 3.6, we used a bin size increasing as 2^k. The distribution of the total, in and out degrees shows the presence of two different regimes, the free-scale property becomes evident for a degree larger than few hundreds. In order to compute the value of the power law exponent γ of Eq. 3.3 we used an ordinary least squares regression, visually identifying the regime zones, and checking for minimal variations of the obtained γ, by progressively widening the degree interval chosen for the regression.

$$p(k) = Ck^{-\gamma} \tag{3.3}$$

The out-degree distribution shows a trend very close to the Barabasi-Albert model, with an exponent $\gamma_{out} = 2.9$, the same value obtained for the generated BA graph. The in-degree distribution decreases with slower speed, having $\gamma_{in} = 2.4$. The total degree distribution exponent has a value of $\gamma_{tot} = 2.6$, intermediate between the in and out distributions.

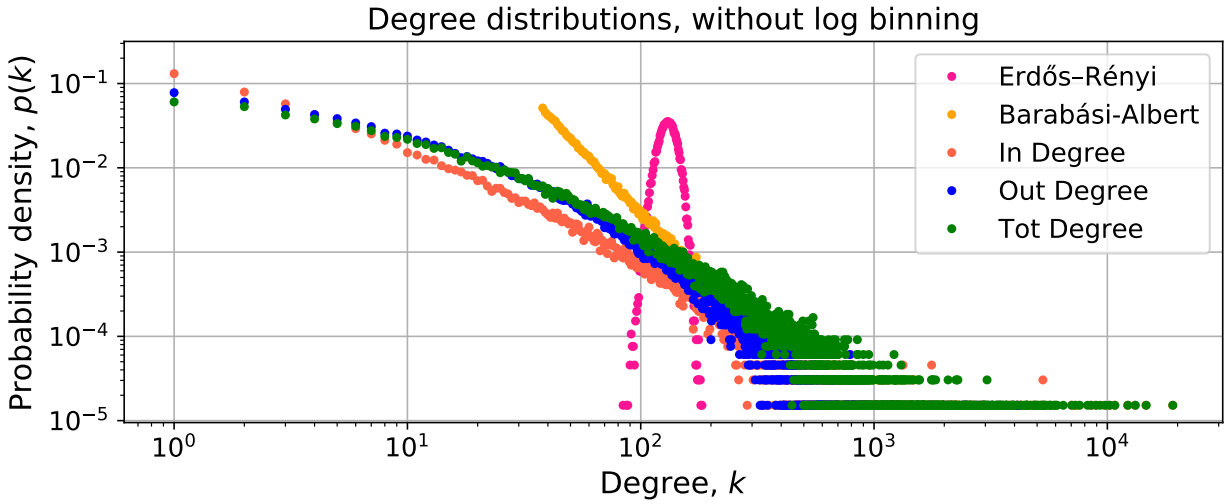


Figure 3.2: Comparison of the degree distributions, plotted without using logarithmic binning.

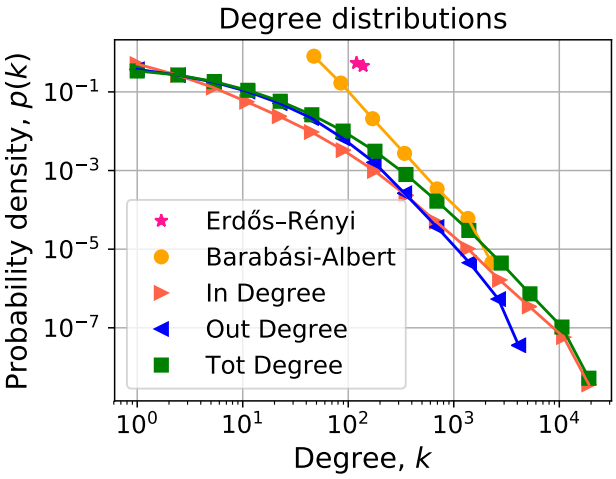


Figure 3.3: Comparison of degree distributions.

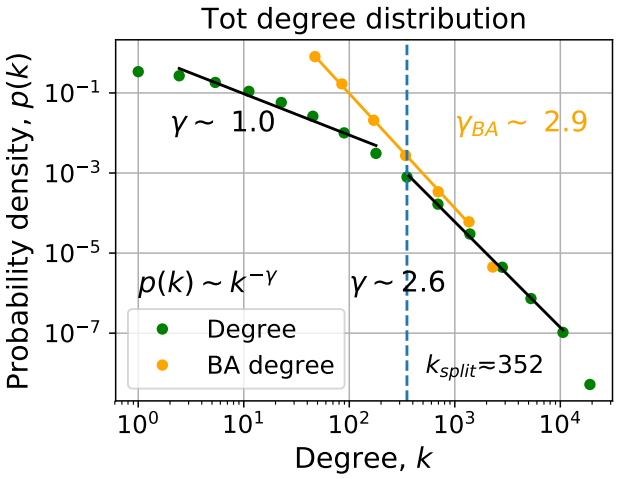


Figure 3.4: Total degree distribution.

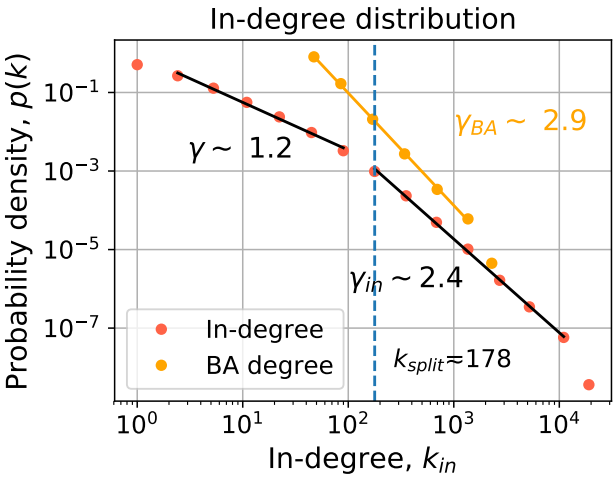


Figure 3.5: In-degree distribution.

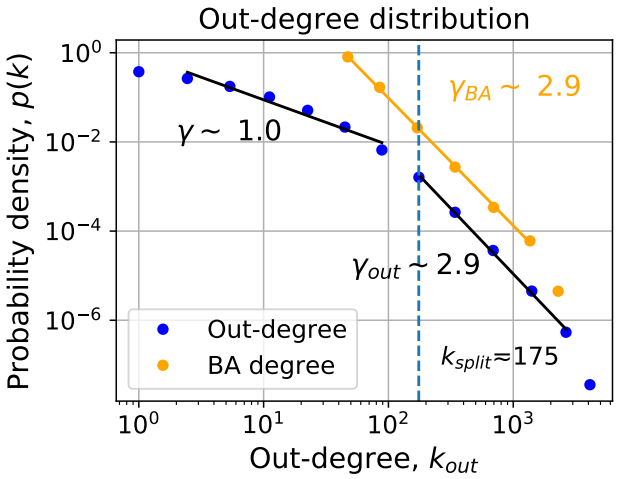


Figure 3.6: Out-degree distribution.

3.3 Path analysis

In order to exactly estimate the average path length $\langle d \rangle$ it would be necessary to compute all the node-node distances of the network. These procedure results infeasible with the computation resources available, as shown in Fig. 3.10. In real networks the path length distribution is quite close to a normal distribution, as shown in [10]. The average path length has then been estimated statistically, random sampling a number n of node pairs, sufficient to achieve a narrow confidence interval for the mean. The assumption of normality of the distribution it is strong, but not necessary. The convergence of the computed mean to the expected value is guaranteed by the central limit theorem with the assumptions that the distances are independent, identically distributed, and with finite variance. The average path length has been estimated by the average of the distances D_i for each sampled node pair, and computing its standard deviation:

$$\langle d \rangle = \frac{\sum D_i}{n} \ , \ \sigma(\langle d \rangle) = \frac{s}{\sqrt{n}} \tag{3.4}$$

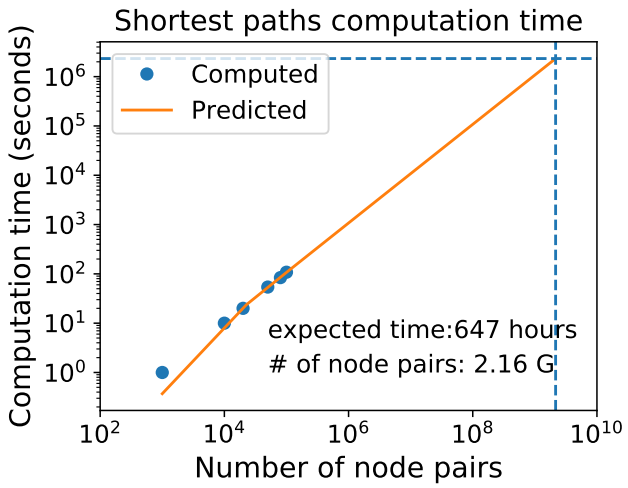


Figure 3.7: Shortest paths computation time by number of pairs

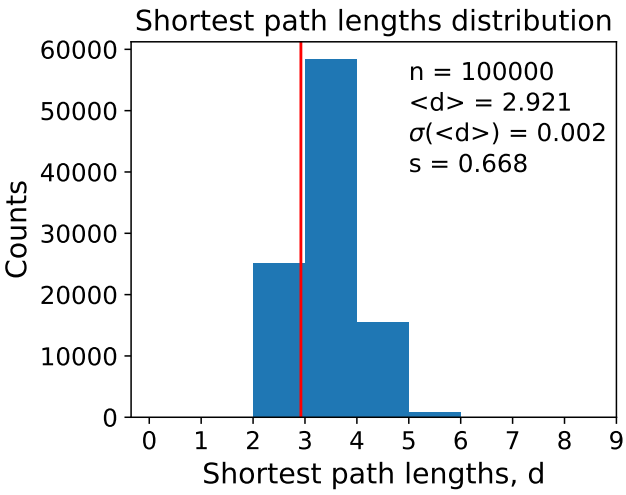


Figure 3.8: Shortest paths distribution

The average shortest path length, computed on the undirected graph, it is equal to 2.92, close to the average distance obtained for both the random graphs, as shown in Fig. 3.9. The shortest path distribution for the original network has a larger dispersion respect to the random ones, there are shortest paths reaching a length equal to 8.

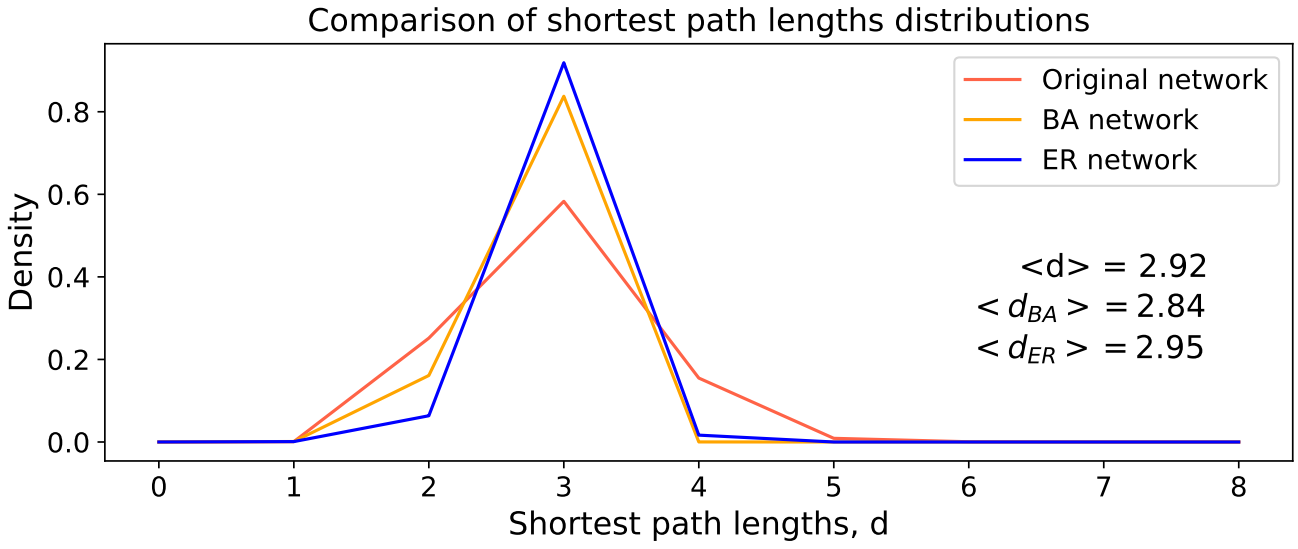


Figure 3.9: Shortest paths distributions comparison between the original undirected graph and the random networks.

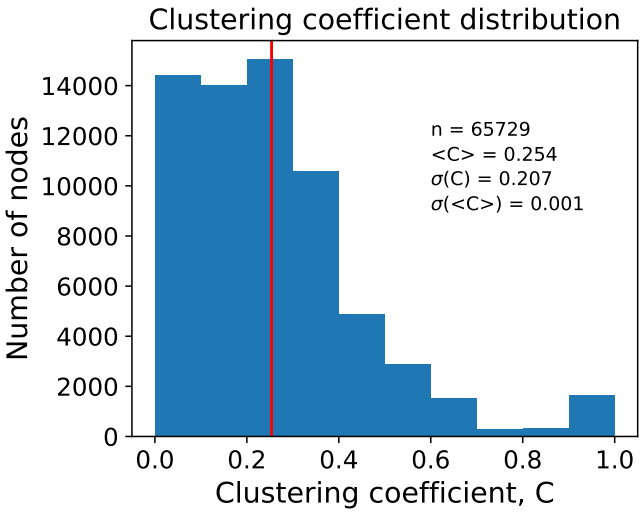


Figure 3.10: Clustering coefficient distribution

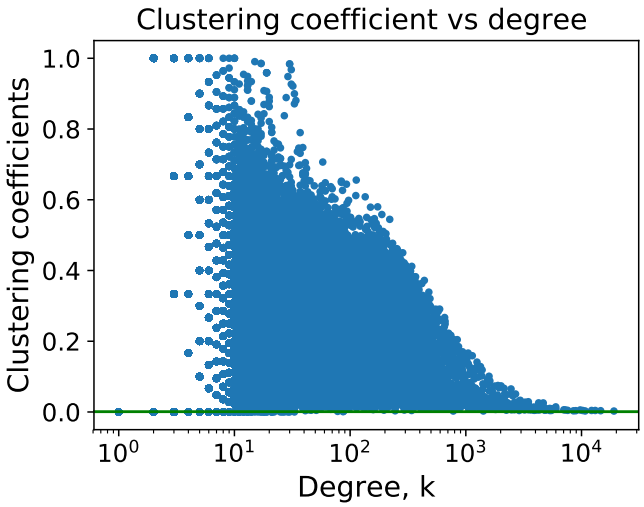


Figure 3.11: Clustering coefficients as function of the degree

3.4 Hubs analysis

The biggest hubs of the crawled social network of tweets authors about the Cambridge Analytica-Facebook scandal are mainly news mass media, as expected. In Fig. 3.12 are represented the 30 biggest hubs (by in-degree) by indicating the in-degree of the crawled network, corresponding to the number of authors following the hub, versus the actual total number of followers on Twitter. The "The New York Times" is the biggest hub, with the maximum number of both in-degree and number of followers. We observe that there is a positive correlation between in-degree and followers, with some variations. In particular, let's take a pair of hubs having similar followers count, such as the "Washington Post" and the "Huffington Post". The "Washington Post" has a larger in-degree than the second. This difference can be interpreted as a larger interest in the scandal from the people following the "Washington Post" respect to the ones following the "Huffington Post". We can define a quantity to measure this interest:

$$\text{Interest} \equiv \frac{\text{in-degree}}{\#\text{followers}} \tag{3.5}$$

This measure represents the fraction of followers that being interested in the scandal had published a tweet about the subject. We can observe for example a large difference in Interest between the two earliest sources, The Guardian and The New York Times, meaning that the followers of the "The Guardian" have relatively interacted much more about the case.

Furthermore we analyzed the correlation of the Interest with the measures in Fig. 3.14, for about 300 of the biggest hubs. The Interest shows an high positive correlation with the Clustering Coefficient, with a value $r = 0.78$, and a mildy negative correlation with the in-degree and the number of followers. The correlation is almost equal to zero respect to the closeness centrality cc.

The correlation between Interest and the Clustering Coefficient C can be interpreted by considering that nodes with higher C have denser connections, that may strengthen the Interest with an high reinforce from the direct neighbours.

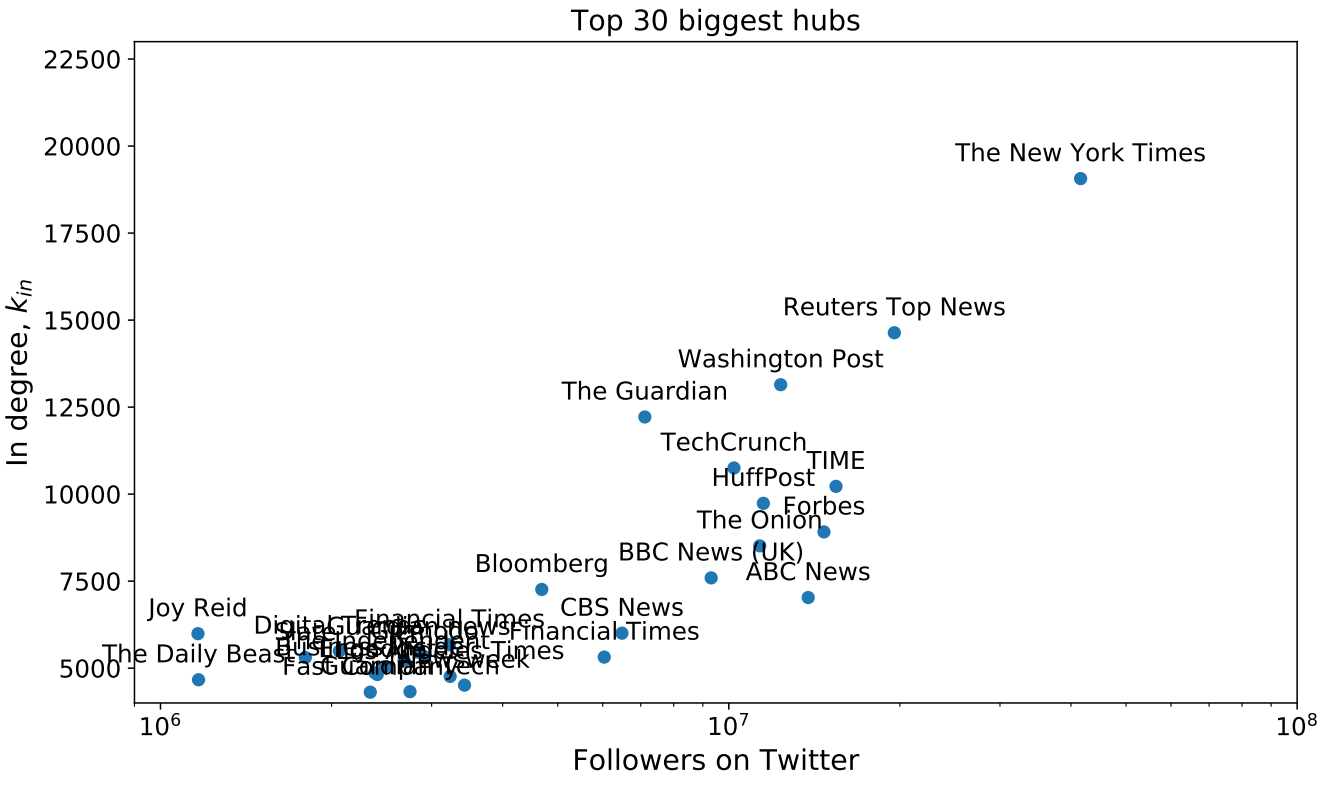


Figure 3.12

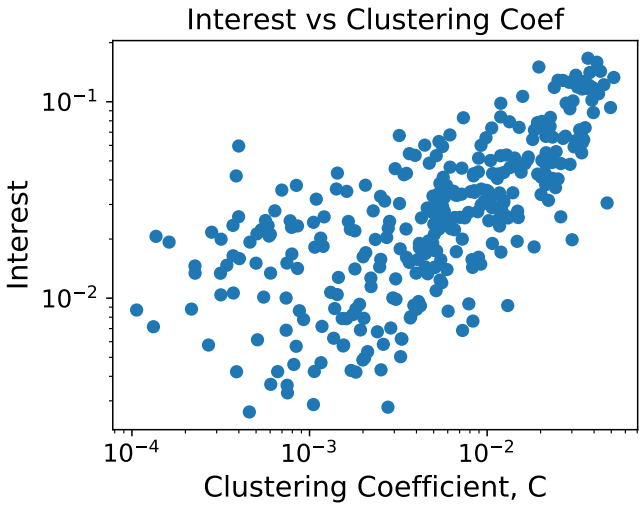


Figure 3.13: Interest versus clustering coefficient for about 300 of the biggest hubs.

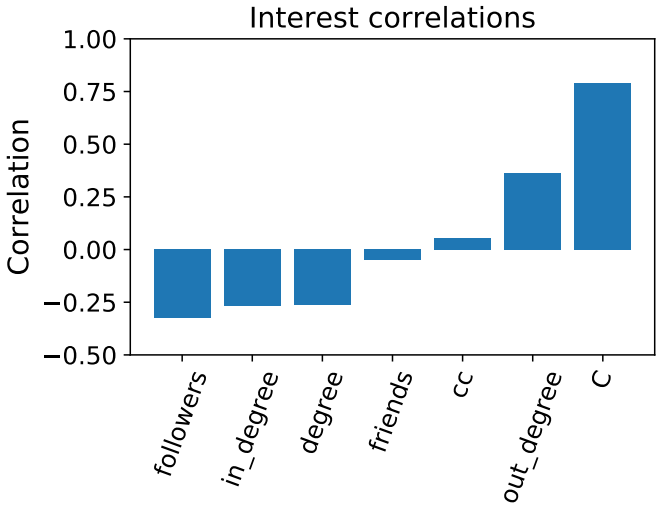


Figure 3.14: Correlations of the Interest with other measures.

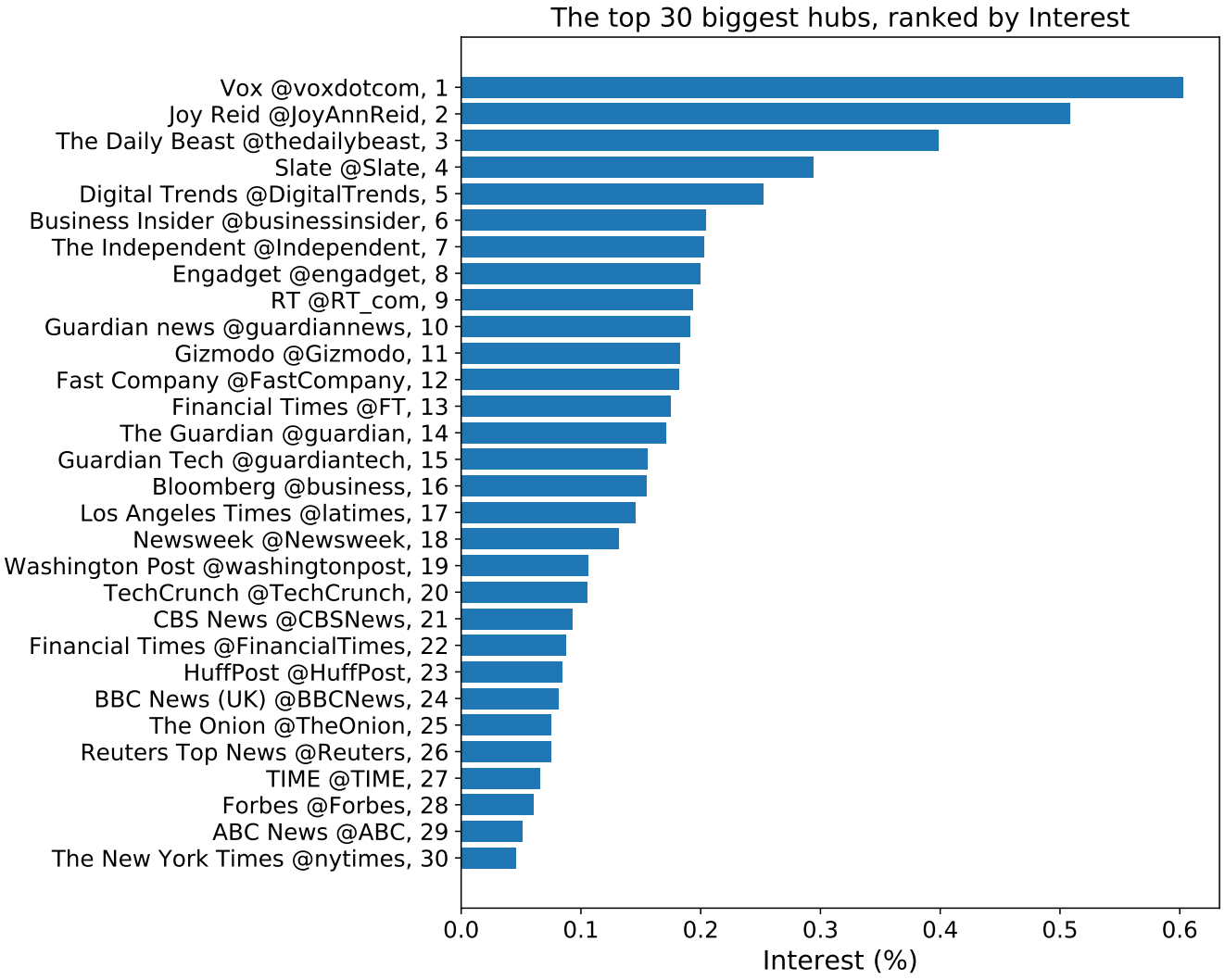


Figure 3.15

3.5 Italian sub network

Some of the analysis on the original graph have been repeated on the Italian sub-network, and are here presented.

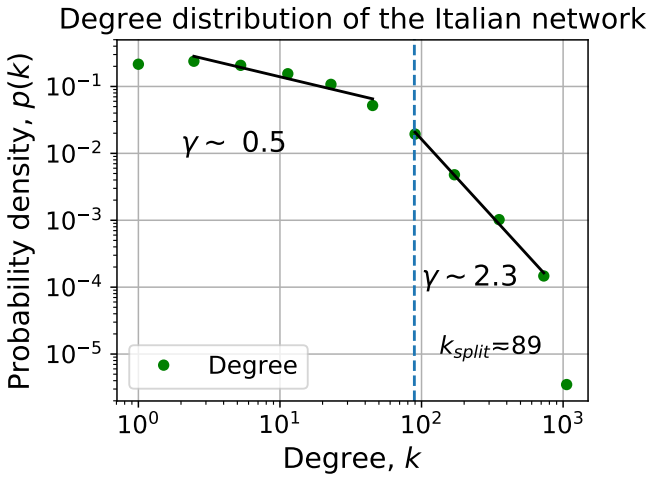


Figure 3.16: Total degree distribution.

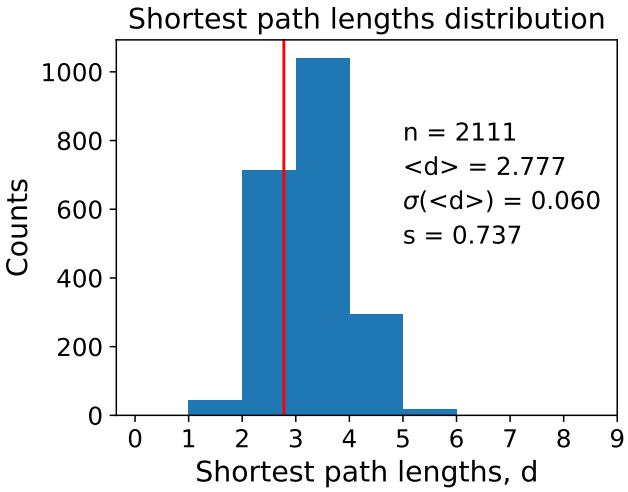


Figure 3.17: Shortest paths distribution.

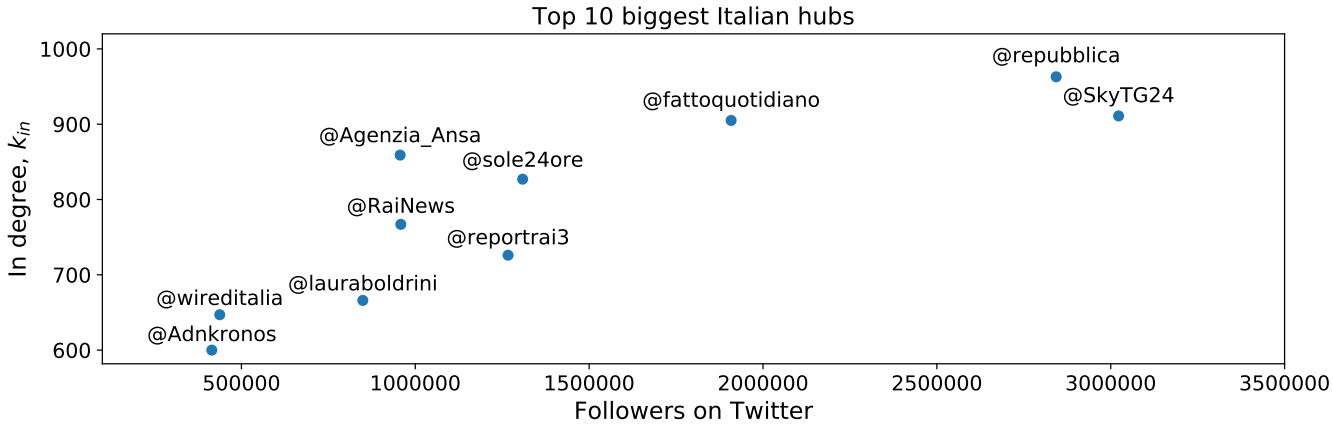


Figure 3.18

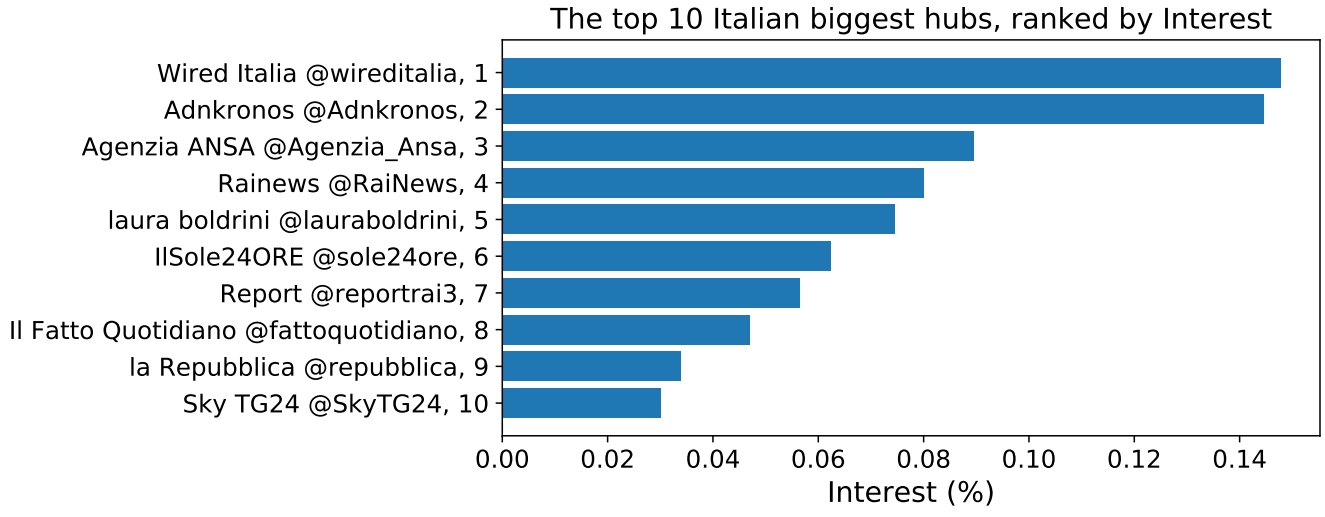


Figure 3.19

4 | Spreading

4.1 Which model for news spreading?

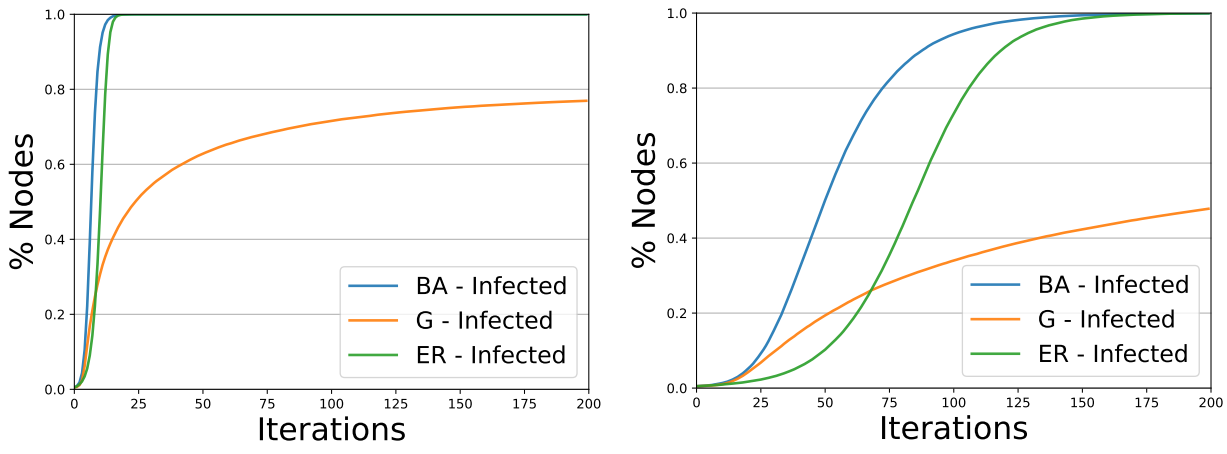
In order to choose a model for the spreading of news or ideas it's necessary to point out what means in this field an infection, and to interpret the related quantities. A person "infected" by a news it's not a person just only reached by the news, but it's a person who partecipates to the spreading by communicating to its neighbours and potentially infect them. The decision to spread the information it's an individual choice, expressing the interest of the person to the news, and indicates the presence of a personal threshold of reaction related to the news or ideas. The threshold can also be influenced by the neighbourhood or the community membership.

In the epidemic models the coefficient β represents the rate of trasmission of the infection and it is constant for the whole population, indicating the dependence on the properties of the pathogen. In the news field β can be interpreted as an intrinsic power of trasmission of the news. This power of trasmission may be associated to the journalistic concept of *newsworthiness*, which includes all the characteristics that make a fact a worthy news. But the expanding phenomenon of *fake news* shows that the speed of diffusion it's not only related to the worthiness of an information. A recent paper published on Science [11] shows how false news on Twitter spread "significantly farther, faster, deeper, and more broadly than the truth in all categories of information". The authors of the paper tried to explain the faster speed of diffusion of the false news by its novelty and the conveying of strongest emotive reactions like surprise or disgust. A news coverage is usually characterized also by a certain amount of time after which the news naturally "dies" out. The SI and SIS epidemic models applied to free-scale networks predict asymptotically scenarios where there is always a finite fraction of infected nodes. From this point of view the SIR model may fit better to reality, predicting the vanishing of the news trasmission. The coefficient μ of the epidemic models, likewise β, it's constant and in this case may represent the intrinsic property of a news to vanish, making the people stopping communicating about it. In summary:

- pathogen or agent: the news or information being spread.
- infected node: a node that is communicating a news, for a certain period of time.
- trasmission rate β: intrinsic power of trasmission of a news/information.
- recover rate μ: intrinsic rate related to a news at which the infected nodes stop communicating about it.
- immunization: the node stops to communicate about the news definitely.
- threshold: personal decision to communicate about the news or not, dependent or not on the neighbours decisions.

In this chapter we'll describe the results we obtained by applying the **SI**, **SIS**, **SIR**, and **Threshold** diffusion models both on the crawled data and on the synthetic graphs (Erdős–Rényi and Barabási–Albert).

4.2 SI model



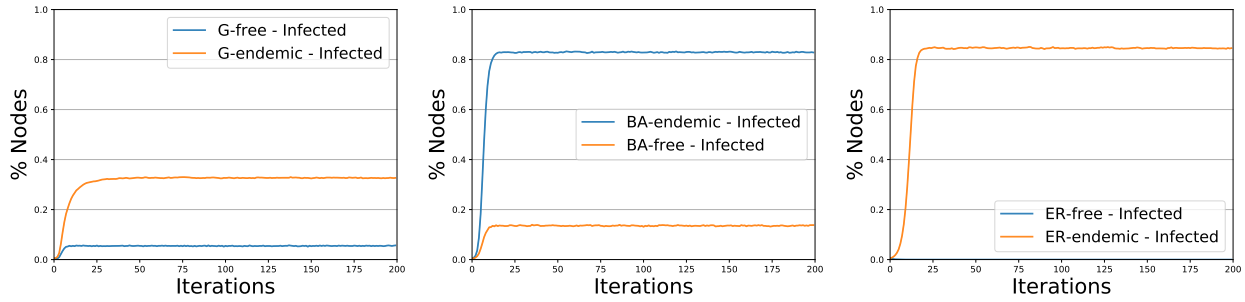
(a) Infection rate $\beta_0 = 0.01$.

(b) Infection rate $\beta_1 = \frac{\beta_0}{10} = 0.001$.

Figure 4.1: Comparison of two trasmission rates in the SI model, for the original network G, the Erdos-Renyi ER and the Barabasi-Albert BA.

For the **Susceptible-Infected** model we've started with a random 0.005% of the total population (3 nodes) of each network being infected, representing the earliest sources of information. In Fig. 4.1 we compare two different trasmission rates, with an order of magnitude of difference: $\beta_0 = 0.01$ for $\beta_1 = 0.001$. The original network asymptotically reach the saturation regime only for the fraction of nodes in the strong connected component.

4.3 SIS model



(a) Original network (b) Barabasi-Albert network (c) Erdos-Renyi network

Figure 4.2: Comparison of the SIS model for the original network G, the Erdos-Renyi ER and the Barabasi-Albert BA.

The introduction of the recovery rate μ in the **Susceptible-Infected-Susceptible** model for networks epidemics provides an epidemic threshold λ_C for the spreading rate λ, dependent on the second order moment of the degree distribution $\langle k^2 \rangle$. For a random network the epidemic threshold given by Eq. 4.1 is finite, and defines two possible asymptotically outcomes, an **endemic state** characterized by a finite fraction of infected individuals, and a completely **disease free** state.

$$\lambda_C(ER) = \frac{1}{\langle k \rangle + 1} \Rightarrow \lambda > \lambda_C : \text{ epidemic state} \tag{4.1}$$

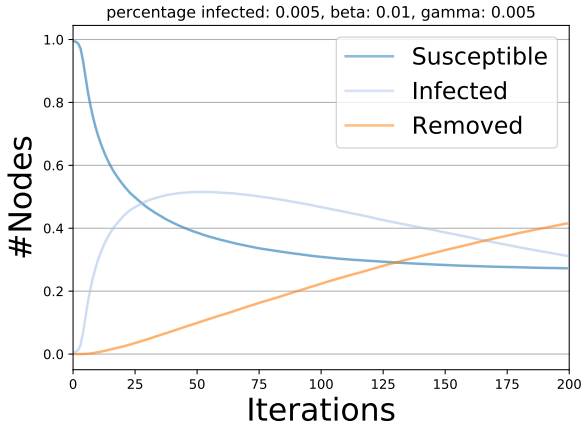
The free-scale networks are characterized by a diverging variance, which means the epidemic threshold tends to vanish, causing a finite fraction of infected individuals also for small λ. We used the random network threshold to choose the recovery rate to simulate both the possible states. We observe in Fig. 4.2 how for the free-scale networks the infected fraction is always finite, below and above the epidemic threshold, while for the random network we observe also the disease free state.

4.4 SIR model

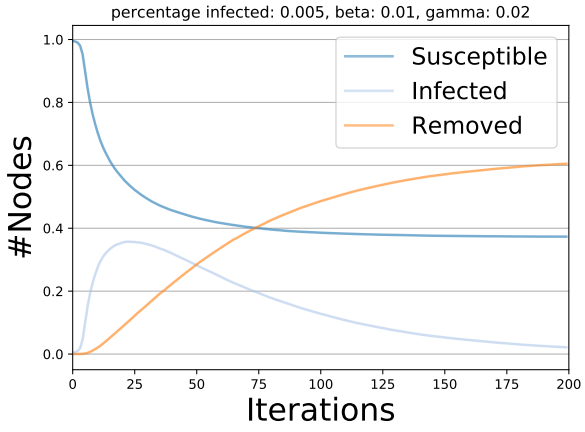
The key characteristic of the **Susceptible-Infected-Recovered** model consists in the possibility of the individuals to recover from the disease and hence to be "removed" from the population instead of returning to the susceptible state. We tested this model either for the case in which μ is smaller than β and the other way around. The different situations are represented in Fig. 4.3, we observe the typical vanishing of the fraction of the infected nodes, after a steep initial rise similar to the one described by the SI model.

4.5 Threshold model

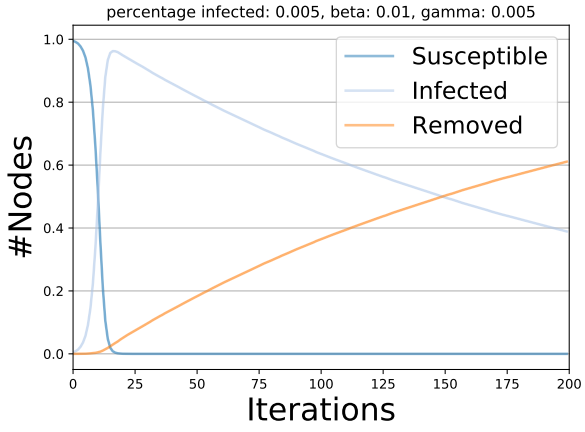
In order to test the **Threshold model** we've chosen a threshold τ equal to 0.10. The diffusion of the infection for this model is represented in Figure 4.4. As we can see, for the original network we have that almost all the nodes become infected within the first 20 model's iterations, due to the fact that the value chosen for the threshold results sufficient for the spreading of the infection. If we change the threshold's value, this time using a value of 0.20, we can observe that the original network become immune to the infection, thanks to its internal structure. We can observe the same immunity in the Erdős–Rényi and Barabási–Albert network for the original threshold's value.



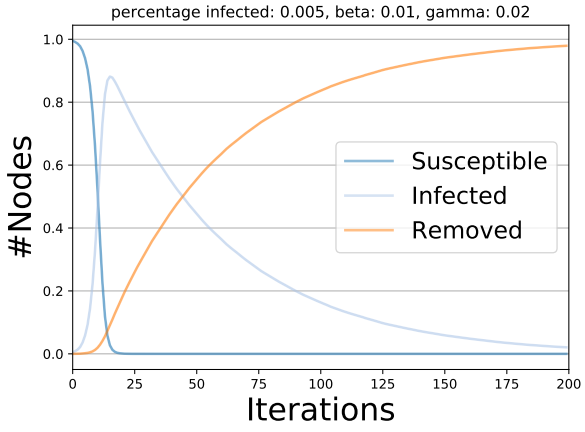
(a) Original network



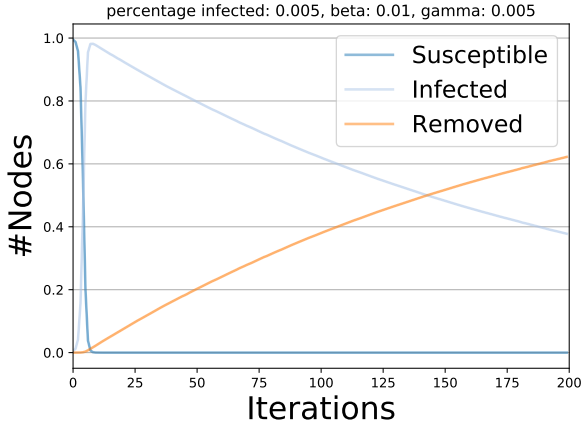
(b) Original network



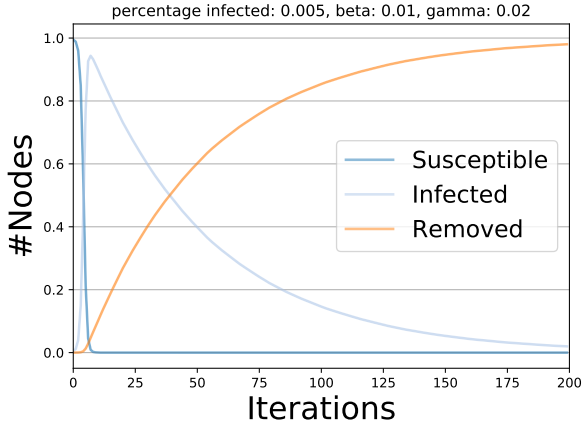
(c) Erdos-Renyi network.



(d) Erdos-Renyi network.

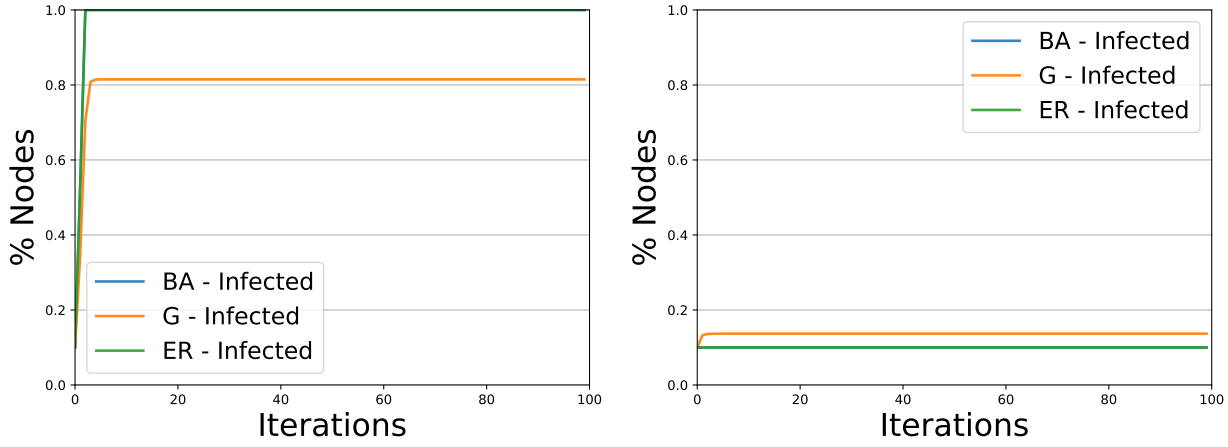


(e) Barabasi-Albert network.



(f) Barabasi-Albert network.

Figure 4.3: SIR model applied to the networks, with two different λ values.



(a) Endemic state with lower threshold, $\tau = 0.1$ (b) Disease-free state with higher threshold, $\tau = 0.2$

Figure 4.4: Threshold model applied to the networks, the Barabasi trend is superimposed by the Erdos-Renyi one.

4.6 The New York Times vs La Repubblica vs Sputnik Italia

Let's assume that the same news it is initially spread by 3 different sources, with very different degree values:

- The New York Times, the biggest hub, with in-degree $k_{in} = 19064$ and #followers=41595294.
- La Repubblica, with $k_{in} = 961$ and #followers=2843075.
- Sputnik Italia, with $k_{in} = 71$ and #followers=6490.

We applied the SIR model obtaining two scenarios, a viral news reaching a large part of the network and a minor news vanishing out after a mildly spread. The scenarios depicted in Fig. 4.5 have been obtained maintaining the same transmission rate β and changing the recovery rate μ. In the viral news scenario the different degree values caused only a modest translation in time of the spread trends, with the same percentages of infected nodes. This result may be explained clearly by the properties of the free-scale networks: the presence of the hubs and the ultra small world property cause a fast spreading from @sputnik_italia to the hubs, who infect immediately a large fraction of the network. In the second scenario, by decreasing the λ ratio below a certain threshold, the news spread by @sputnik_italia nodes probably did not reach the hubs and the infection did not propagate.

The two presented scenarios can explain the fact that for example *fake news* can also begin spreading from peripheral, small nodes and propagates virally over a network. The difference between the two scenarios is carried by the intrinsic transmission properties of the news, independently from the properties of the carrier: the message it is more important than the messenger.

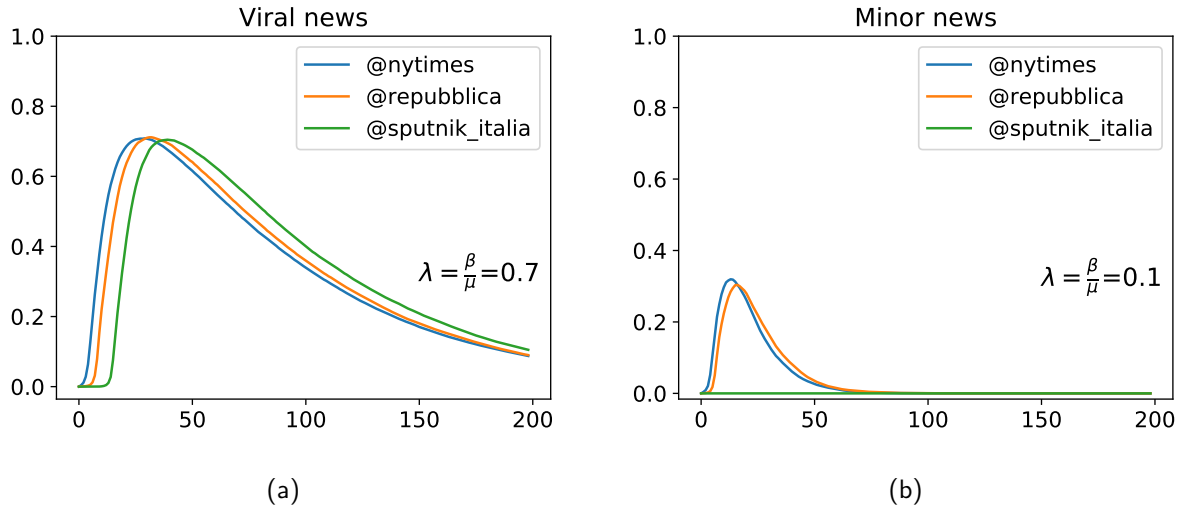


Figure 4.5: Viral news spreading: infected percentages trends.

5 | Communities discovery

In this chapter we'll provide the results obtained by applying **K-Clique**, **Label Propagation**, **Louvain**, **Girvan-Newman** and **Demon** to a sample of 2000 nodes taken from the original network. We've chosen to sample the crawled data in order to ease the application of the various algorithms. Each partition is evaluated by applying an implementation of the scoring functions listed in [12], and, for each algorithm, the results are represented in a table. Together with the results of the scoring functions are also provided the total number of communities discoverd (Communities), the number of nodes in the smallest/biggest community (Smallest/Biggest), the hashtags utilized in the biggest community (Tags) and finally the languages of the users in the biggest community (Langs). Finally it is also provided an evaluation for the subset of the original network composed only by italian users.

5.1 K-Clique

We have chosen to apply the **K-Clique** algorithm, described in [13], to the sample using three different values for k: 3, 4 and 5, respectively. The results are represented in Table 5.1. As we can see, even if the result is not so good, the best partition is obtained by using k eguals to 3, which returns a low modularity partition composed by low degree nodes. It is interesting to note that the biggest partitions returned by the first application of the algorithm is composed by english and deutsch speaking users, while the other iterations returns community composed only by english speaking users.

K	Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
3	29	51	3	0.14	0.62	0.20	2.72	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, Deutsch
4	13	14	4	0.020	0.68	0.21	4.23	cambridgeanalytica, deletefacebook, facebook	English
5	5	12	5	0.011	0.63	0.23	4.83	cambridgeanalytica, deletefacebook, facebook	English

Table 5.1: Evaluation of the partitions obtained by the application of the K-Clique algorithm.

5.2 Label Propagation

In Table 5.2 are represented the results of the application of the **Label Propagation** algorithm, described in [14]. According to the modularity score the partition provided by this algorithm represent a good subdivision of the original network, even if it is composed for the vast majority by small communities, as suggested by the high number of communities and the low value for the Average Node Degree score. Since this partition is composed by an high number of communities, the biggest community aggregates a vast variety of languages.

Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
1278	136	1	0.68	0.28	0.19	1.38	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various

Table 5.2: Evaluation of the partition obtained by the application of the Label Propagaion algorithm.

5.3 Louvain

The application of the **Louvain** algorithm, described in [15], along with the iteration of the Label Propagation algorithm, returns the best partition among all the partitions returned by the other algorithms. The results of its application are represented in Table 5.3. As for the Label Propagation algorithm, this partition also is composed by an high number of small communities, and the biggest community is composed by users who speaks various languages.

Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
1193	133	1	0.76	0.042	0.17	1.60	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various

Table 5.3: Evaluation of the partition obtained by the application of the Louvain algorithm.

5.4 Girvan-Newman

For the **Girvan-Newman** algorithm, described in [16], we've decided to record the results of 5 iterations over the sample network. The results are represented in Table 5.4. As you can see, the first three iterations returns very poor partitions, with low modularity scores, due to the fact that the edges with the highest betweenness

centrality selected in the various iterations doesn't provide a good grade of separation among the nodes of the network. With the fourth and fifth iterations, there is a consistent improvement either in the modularity score and in the other measures.

Iteration	Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
1	1181	703	1	0.17	0.00098	0.22	1.23	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various
2	1182	659	1	0.24	0.0019	0.21	1.27	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various
3	1183	588	1	0.38	0.0048	0.21	1.35	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various
4	1184	575	1	0.40	0.0081	0.20	1.34	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	Various
5	1185	458	1	0.57	0.011	0.20	1.40	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	Various

Table 5.4: Evaluation of the partition obtained by the application of the Girvan-Newman algorithm.

In general, the partitions returned by the five iterations of the algorithms are all composed by an high number of small communities, with the biggest community growing smaller iteration after iteration. This kind of fragmentation, as seen before, for every iteration produces a biggest community with diffent types of users.

5.5 Demon

Finally in Table 5.5 we provide the results of the application of the **Demon** algorithm, described in [17], that we tested for five different values of ϵ, 0.10, 0.25, 0.50, 0.75 and 0.90, respectively. In general, the five partitions are not so good from the point of view of the modularity score, with the third application of the algorithm beign the best. Contrary to the results obtained by the application of the other algorithms, the partitions for the Demon algorithm are all composed by a small number of communities.

Epsilon	Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
0.10	10	147	4	0.07	0.46	0.082	4.41	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	Various
0.25	11	63	4	0.095	0.44	0.094	4.31	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, Deutsch
0.50	21	43	4	0.11	0.57	0.10	4.15	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, Deutsch
0.75	39	25	4	0.068	0.62	0.12	3.96	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, Deutsch
0.90	89	24	4	0.071	0.68	0.15	3.65	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, Deutsch

Table 5.5: Evaluation of the partition obtained by the application of the Demon algorithm.

5.6 Comparisons

In this section of the chapter we compare the algorithms used so far by confronting the best instances among the iterations provided in the past sections. The comparisons are performed by using the NF1 score, as described in [18].

A1	A2	F1 mean	Ground Truth Communities	Identified Communities	Community Ratio	Ground Truth Matched	Node Coverage	NF1
K-Clique 3	Label Propagation	0.44	714.0	21.0	0.029	0.022	0.10	0.0076
K-Clique 3	Louvain	0.32	683.0	21.0	0.031	0.016	0.10	0.0027
K-Clique 3	Girvan-Newman 5	0.25	677.0	21.0	0.031	0.012	0.10	0.0011
K-Clique 3	Demon 0.50	0.94	9.0	21.0	2.33	0.78	1.44	0.24
Label Propagation	Louvain	0.94	1193.0	1278.0	1.07	1.0	1.0	0.87
Label Propagation	Girvan-Newman 5	0.93	1185.0	1278.0	1.08	1.0	1.0	0.86
Label Propagation	Demon 0.50	0.26	21.0	1278.0	60.86	0.76	10.24	0.0025
Louvain	Girvan-Newman 5	0.99	1185.0	1193.0	1.0068	1.0	1.0	0.99
Louvain	Demon 0.50	0.32	21.0	1193.0	56.81	0.62	10.24	0.0021
Girvan-Newman 5	Demon 0.50	0.36	21.0	1185.0	56.43	0.24	10.24	0.00036

Table 5.6: Comparisons among the best iterations of the algorithms utilized in this chapter.

In Table 5.6 we can see the comparisons among the best iterations of the algorithms utilized during the community discovery phase. As we can see by the results, the best comparisons are the ones between K-Clique and Demon , with, respectively, $k = 3$ and $\epsilon = 0.50$, between Label Propagation and Louvain/Girvan-Newman (fifth iteration) and finally between Louvain and Girvan-Newman (fifth iteration).

5.7 The italian subgraph

In order to better understand the composition of our network, we selected among the nodes only the ones composed by **italian users**, and on this subgraph we've applied the same algorithms described in the past

sections. The obtained partitions make clear to us that the italian users inside our network aren't disposed into a community-like shape. The best obtained result is the one returned by the Louvain algorithm, which return a modularity score of 0.26 and a partition composed by 6 communities, the biggest and the smallest being composed by 817 and 2 users, respectively. Being the partition composed by only 6 communities, the average node degree is bigger (≈ 17) that the ones returned by the other evaluations.

6 | Network robustness

In this chapter we'll provide some results about the **robustness** and **attack tolerance** of our network. Thaking as reference the concepts described in [9], we'll define the **critical threshold** of our network, and we'll test its robustness against attacks conducted following a random nodes' selection or one based on decreasing degree centrality. Finally we'll test the **Failure Propagation Model**, following our implementation of the model, on our network.

6.1 Critical threshold

As described in [9], we have obtained the **critical threshold** representing the fraction of the nodes that must be removed to break apart our network. This fraction, represented by f_c, is obtained by the following formula:

$$f_c = 1 - \frac{1}{\frac{\gamma-2}{3-\gamma} k_{min}^{\gamma-2} k_{max}^{3-\gamma} \ - \ 1} = 1 - \frac{1}{1.50 * 1^{0.6} * 19073^{0.4} \ - \ 1} = 0.99$$

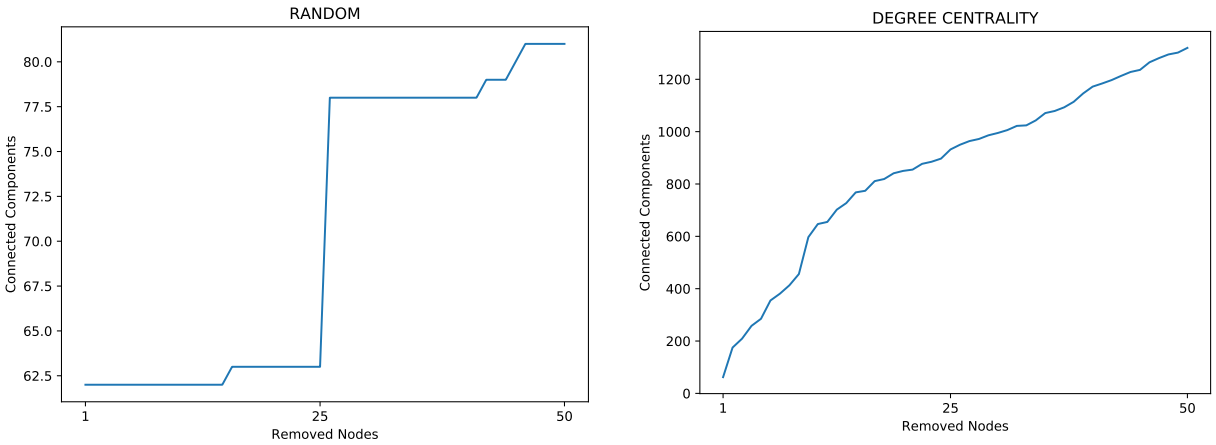
which, remembering that the γ for our scale-free network corresponds to 2.6 and that k_{min} and k_{max} are eguals to 1 and 19073 respectively, tells us that, in order to break apart our network it is mandatory to remove the 99% of the nodes. Keeping in mind that our network is, in fact, a finite network, we can adjust the obtained result by utilizing the following formula, still in [9]:

$$f_c \approx 1 - \frac{C}{N^{\frac{3-\gamma}{\gamma-1}}} \approx 1.00$$

where $C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}}$ is a constant and N represents the number of nodes of the network. As we can see, this new approximation tells us that in order to break apart our network the totality of its nodes must be removed.

6.2 Simulation of an attack

In order to validate the results obtained in Section 6.1, here we simulate an attack to our network. We've chosen to simulate the remotion of 50 nodes from the network following two distinct criterions: **random selection** and **degree centraility** (decreasing order). For every criterion we've monitored the fragmentation of the connected components.



(a) (b)

Figure 6.1: In Figure 6.1a we can see the fragmentation of the connected components during the remotion based on a randomic choice, while in Figure 6.1b we can see the same fragmentation, but this time based on decreasing degree centraility.

As we can see, for the randomic choice of the nodes to be removed, the structure of the network is barely altered. After the random remotion of 50 nodes, the original 62 connected components became slightly more than 80. For the remotion of the nodes based on decreasing degree centrality there is, as expected, a different situation. This kind of criterion guarantees that the original structure of the network is broken apart more easily, because the original network's hubs are removed one by one in decreasing order.

6.3 Simulation of a Failure Propagation Model

To test more the robustness of our network, we've written the code in order to implement (and test) the **Failure Propagation Model**, as described in [9]. In Figure 6.2 you can see some iterations of the model in which we used different values for the φ parameter.

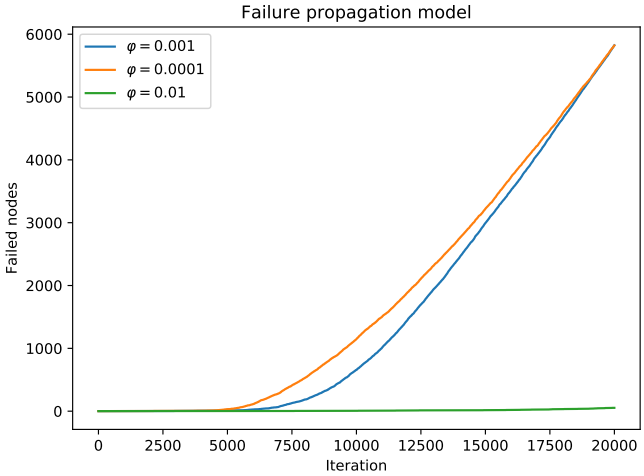


Figure 6.2: The test for the Failure Propagation Model was conducted over 20000 iterations.

We can see that, as expected, the network doesn't accuse the failure propagation for φ eguals to 0.01, in which less than 30 nodes failed. For the smaller values of φ we can see that the situation is different, with a greater amount of nodes which fail over the iterations.

7 | Summary

References

[1] The Guardian. *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. 2018.

[2] New York Times. *How Trump Consultants Exploited the Facebook Data of Millions*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html. 2018.

[3] New York Times. *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. 2018.

[4] Channel 4. *Exposed: Undercover secrets of Trump's data firm*. https://www.channel4.com/news/exposed-undercover-secrets-of-donald-trump-data-firm-cambridge-analytica. 2018.

[5] Ars Technica. *Facebook scraped call, text message data for years from Android phones*. https://arstechnica.com/information-technology/2018/03/facebook-scraped-call-text-message-data-for-years-from-android-phones/?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axiosam&stream=top-stories. 2018.

[6] EUR-lex. *General Data Protection Regulation (GDPR) (EU) 2016/679*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L:2016:119:FULL&from=EN. 2016.

[7] The Guardian. *A radical proposal to keep your personal data safe. Richard Stallman*. https://www.theguardian.com/commentisfree/2018/apr/03/facebook-abusing-data-law-privacy-big-tech-surveillance. 2018.

[8] SoBigData.eu. *First Aid For Data Scientist*. http://fair.sobigdata.eu/moodle/. 2018.

[9] Albert-László Barabási and Márton Pósfai. *Network Science*. http://networksciencebook.com/. 2016.

[10] Qi Ye, Bin Wu, and Bai Wang. "Distance Distribution and Average Shortest Path Length Estimation in Real-world Networks". In: *Proceedings of the 6th International Conference on Advanced Data Mining and Applications: Part I*. ADMA'10. Chongqing, China: Springer-Verlag, 2010, pp. 322–333. ISBN: 3-642-17315-2, 978-3-642-17315-8. URL: http://dl.acm.org/citation.cfm?id=1947599.1947633.

[11] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151. ISSN: 0036-8075. DOI: 10.1126/science.aap9559. eprint: http://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: http://science.sciencemag.org/content/359/6380/1146.

[12] Yang Jaewon and Leskovec Jure. *Defining and Evaluating Network Communities based on Ground-truth*. https://link.springer.com/article/10.1007/s10115-013-0693-z. 2015.

[13] Gergely Palla et al. *Uncovering the overlapping community structure of complex networks in nature and society*. http://dx.doi.org/10.1038/nature03607. 2005.

[14] G. Cordasco and L. Gargano. *Community detection via semi-synchronous label propagation algorithms*. https://ieeexplore.ieee.org/abstract/document/5730298/citations. 2010.

[15] Vincent D Blondel et al. *Fast unfolding of communities in large networks*. http://stacks.iop.org/1742-5468/2008/i=10/a=P10008. 2008.

[16] M. Girvan and M. E. J. Newman. *Community structure in social and biological networks*. http://www.pnas.org/content/99/12/7821. 2002.

[17] Michele Coscia et al. *DEMON: a Local-First Discovery Method for Overlapping Communities*. http://arxiv.org/abs/1206.0629. 2012.

[18] Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo. *A Novel Approach to Evaluate Community Detection Algorithms on Ground Truth*. https://doi.org/10.1007/978-3-319-30569-1_10. 2016.