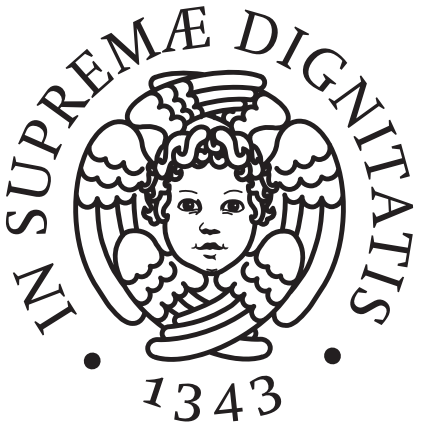

Università di Pisa

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 Cambridge Analytica datascientist, about its leaders.

Contents

1 The case story

On Saturday 17 of March 2018, the newspapers The Observer and The New York Times 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]. REF OBSERVER

The whistleblower Christopher Wylie, datascientist and former director of research at Cambridge Analytica revealed... Cambridge Analytica described itself as a company providing consumer research, targeted advertising and other data-related services to both political and corporate clients.

What, Where, Who, Why, Where ?

Timeline da sistemare: [2]

- March 17, 2018: 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, 2018: 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.
- March 20, 2018: 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.
- to be continued...

2 | Building the network

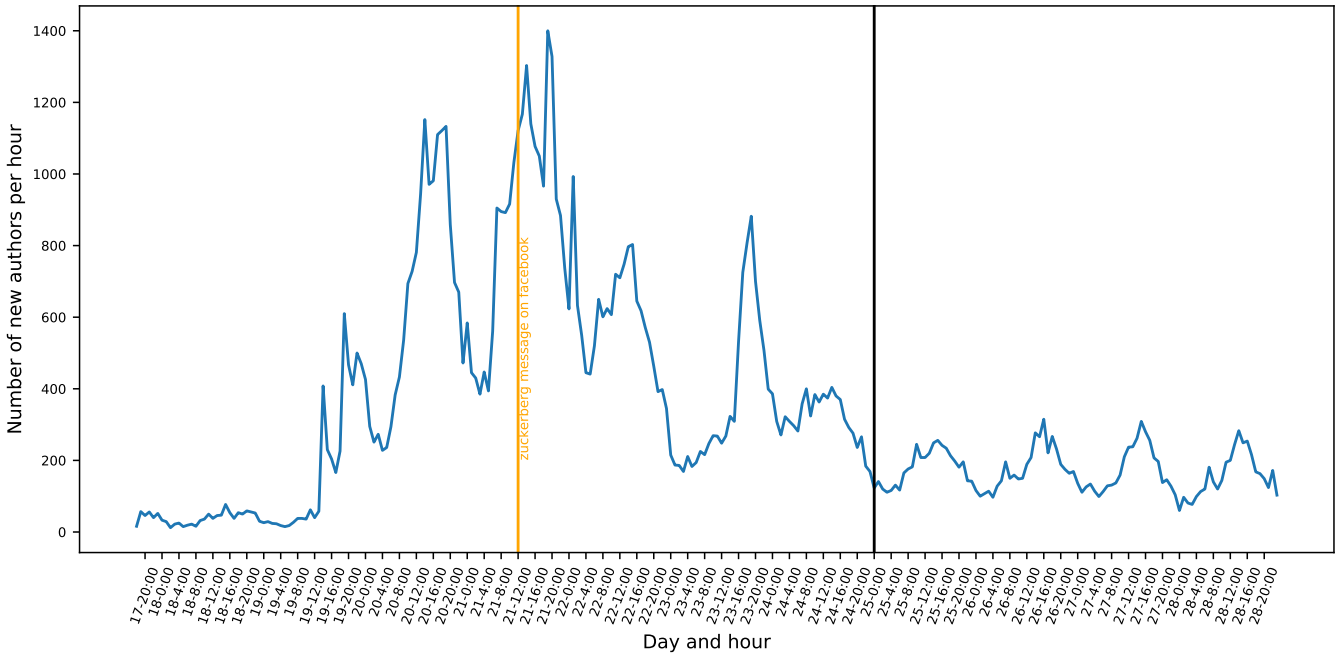

Figure 2.1: New authors time history

3 | Network properties

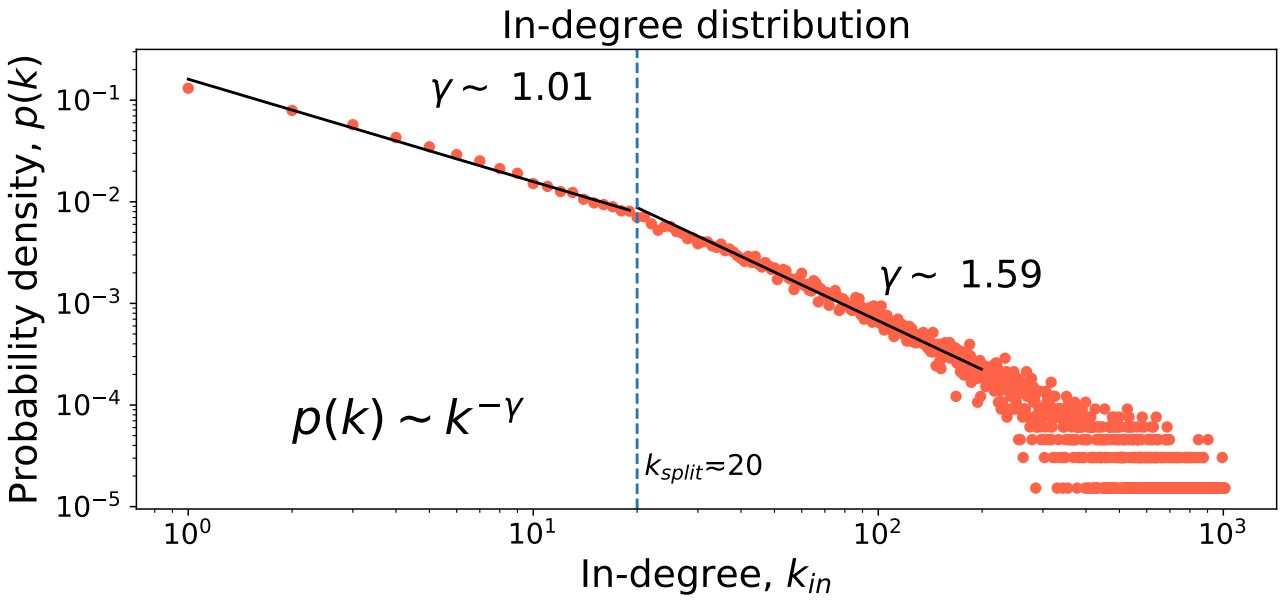

Figure 3.1: New authors time history

4 | Network dynamics

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 1000 nodes taken from the original network. We've chosen to sample the crawled data in order to ease the application of the various algorithms. For each application of the algorithms we provide a table containing the measures for evaluating the obtained partitions (Modularity, Conductance, IED and AND), 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).

5.1 K-Clique

We have chosen to apply the **K-Clique** algorithm 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, there are no communities resulting from the application of 5-Clique.

K	Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
3	8	12	3	0.086	0.53	0.17	2.97	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English
4	5	4	4	0.0092	0.63	0.25	3.0	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. Together with the Louvain algorithm, it returns the best partition, as we can see by the modularity and conductance scores.

Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
757	46	1	0.66	0.24	0.18	1.42	zuckerberg, cambridgeanalytica, deletefacebook, facebook ,facebookgate	English

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 returns the best partition among all the partitions returned by the other algorithms. The results of its application are represented in Table 5.3.

Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs
731	34	1	0.75	0.070	0.14	1.62	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	English, French, Deutsch, Arabic

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

5.4 Girvan-Newman

For the **Girvan-Newman** algorithm, 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 iteration returns a very poor partition, with a low modularity score, due to the fact that the edge with the highest betweenness centrality in the starting sample network doesn't provide a good grade of separation among the nodes of the network. With the second iteration, and the ones after that, 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	720	234	1	0.19	0.0016	0.21	1.24	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	English, French, Deutsch, Arabic, Spanish, Italian, Portoguese, Hindi, Finnish, Hungarian
2	721	117	1	0.55	0.0077	0.20	1.32	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, French, Deutsch, Hindi, Finnish, Spanish
3	722	117	1	0.59	0.016	0.19	1.36	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	English, French, Deutsch, Arabic, Italian, Portoguese, Hungarian
4	723	109	1	0.61	0.018	0.18	1.42	cambridgeanalytica, deletefacebook, privacy, zuckerberg, facebook, facebookgate	English, French, Deutsch, Arabic, Italian, Portoguese, Hungarian
5	724	96	1	0.65	0.023	0.18	1.49	zuckerberg, cambridgeanalytica, deletefacebook, facebook, facebookgate	English, French, Deutsch, Italian, Hindi

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

5.5 Demon

Epsilon	Communities	Biggest	Smallest	Modularity	Conductance	IED	AND	Tags	Langs

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

6 | Spreading

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) generated from the original one. In each section, a comparison between the three networks will be provided along with some details on the implementation of the tests of every model.

6.1 SI model

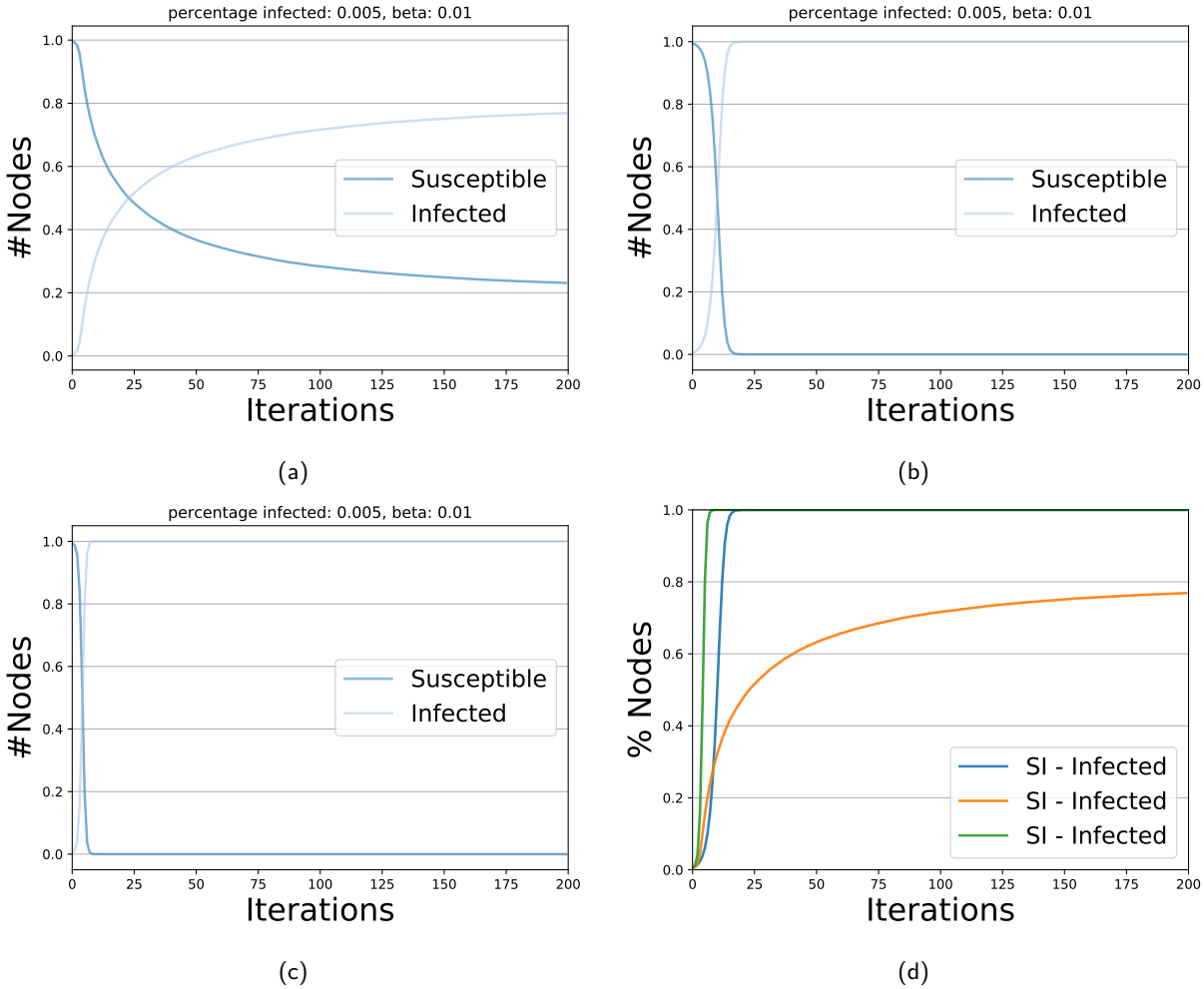

Figure 6.1: In Figure 6.1a we can see the diffusion graph for the original network, while in Figure 6.1b and in Figure 6.1c we can see the diffusion graph for the Erdős–Rényi and Barabási–Albert networks, respectively. In Figure 6.1d we can see a comparison between the infection rate of the three networks.

For the **Susceptible-Infected** model we've started with a 0.005% of the total population (3 nodes) of each network being infected, and we've choosed a value of 0.01 for the infection rate β. As you can see from Figure 6.1, the original network is the only one that doesn't reach the saturation regime, while the other networks reach it within the first 25 iterations of the model. This is due to the fact that both the Erdős–Rényi and the Barabási–Albert network are extremely connected, hence it is more easy for the infection to spread among the nodes.

6.2 SIS model

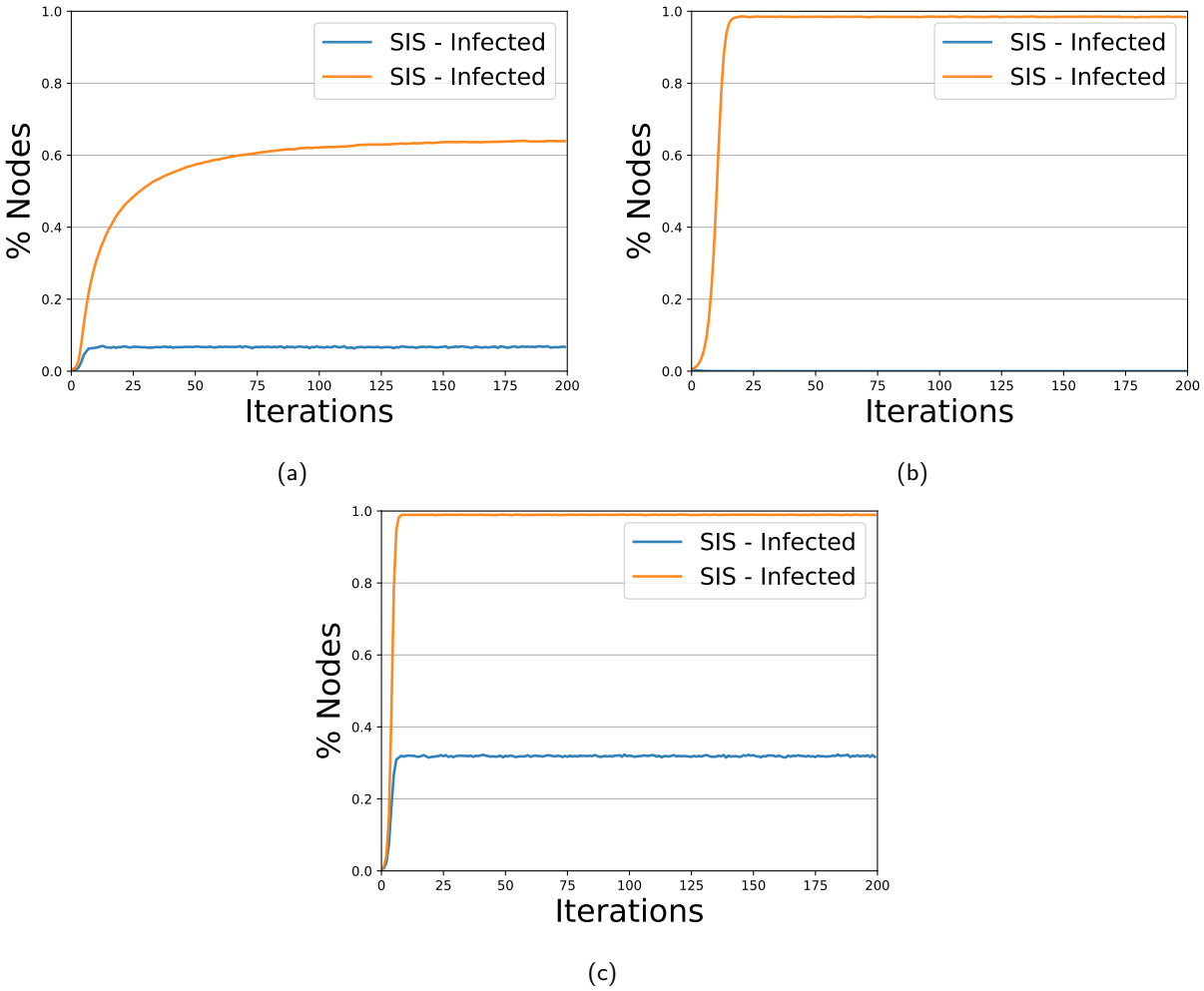

Figure 6.2: In Figure 6.2a we can see the comparison between the endemic state, in orange, and the disease free state, in blue, for the original network. The same comparison can be observed for the Erdős–Rényi and the Barabási–Albert network, respectively, in Figure 6.2b and 6.2c

For the **Susceptible-Infected-Susceptible** model, thanks to the introduction of the recovery rate μ, we can model two possible outcomes for the epidemic: the **endemic state**, characterized by a low recovery rate and by the fraction of infected individuals that follows a logistic curve similar to the one observed for the SI model, for which $\mu < \beta\langle k\rangle$, and the **disease free** state, characterized by a sufficiently high recovery rate, for which $\mu > \beta\langle k\rangle$. A comparison between this two states is represented for every network in Figure 6.2.

6.3 SIR model

The key characteristic of the **Susceptible-Infected-Recovered** model consist in introducing the probability γ for the individuals to recover from the disease and hence to be "removed" from the population instead of returning to the susceptible state. We have choosen to test this model either for the case in which γ is smaller than β and the other way around. The graphs representing this different situations for all the three networks are visible in Figure 6.3.

6.4 Threshold model

Finally we describe the application of the **Threshold model** both on the original network and the synthetic ones. In order to test this model we've choosen to apply a threshold τ eguals to 0.10, the diffusion of the infection for this model is represented in Figure 6.4. 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 choosen for the threshold results to be sufficient for the

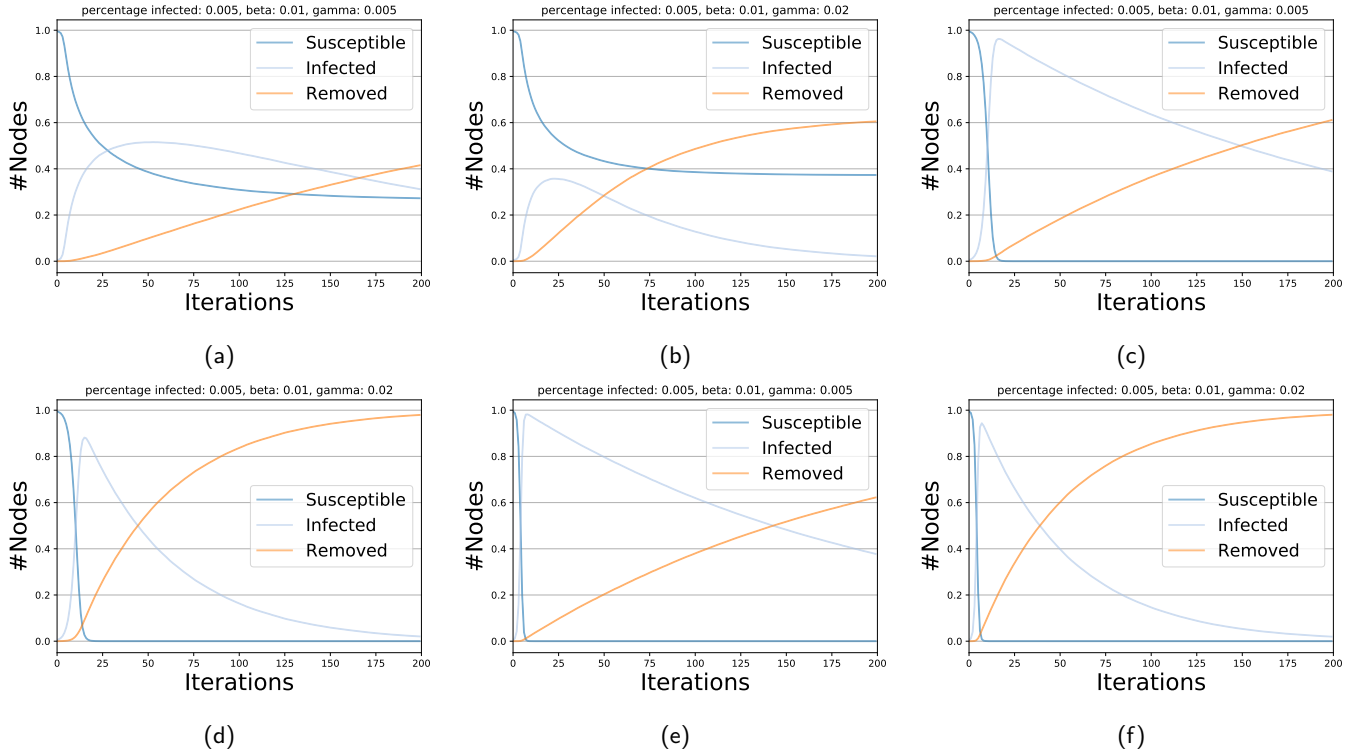

Figure 6.3: In Figure 6.3a and 6.3b we can see the representation of the diffusion on the original network both for the case in which γ is smaller than β and the other way around. The same kind of representation is plotted for the Erdős–Rényi network in Figure 6.3c and 6.3d and for the Barabási–Albert network in Figure 6.3e and 6.3f.

spreading of the infection. If we change the threshold's value, this time using 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.

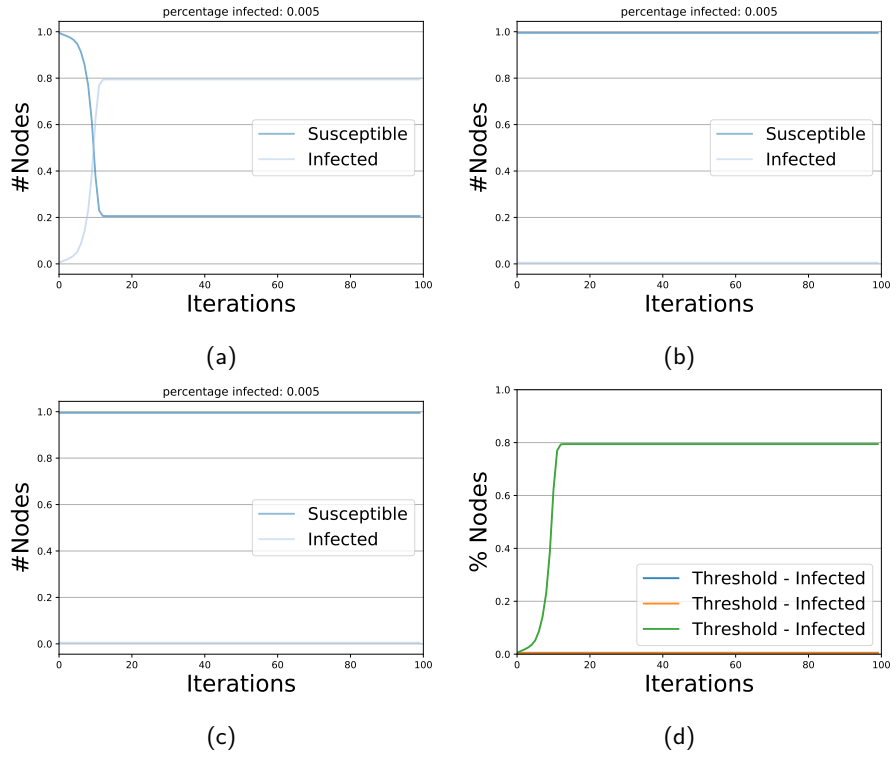

Figure 6.4: In Figure 6.4a is represented the diffusion of the infection for the original network, while in Figure 6.4b and 6.4c are represented the cases for the Erdős–Rényi and the Barabási–Albert network, respectively. A comparison between the three networks is represented in Figure 6.4d.

7 | Summary

References

[1] 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. [Online; accessed 19-May-2018]. 2018.

[2] 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. [Online; accessed 19-May-2018]. 2018.