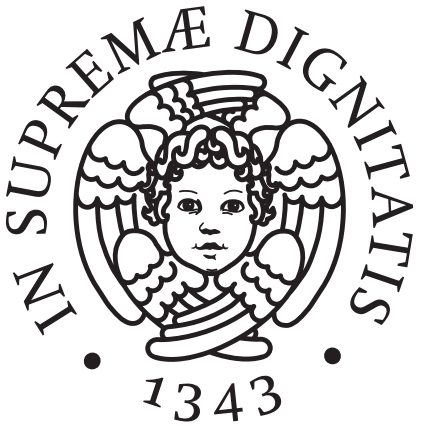

Università di Pisa

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 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

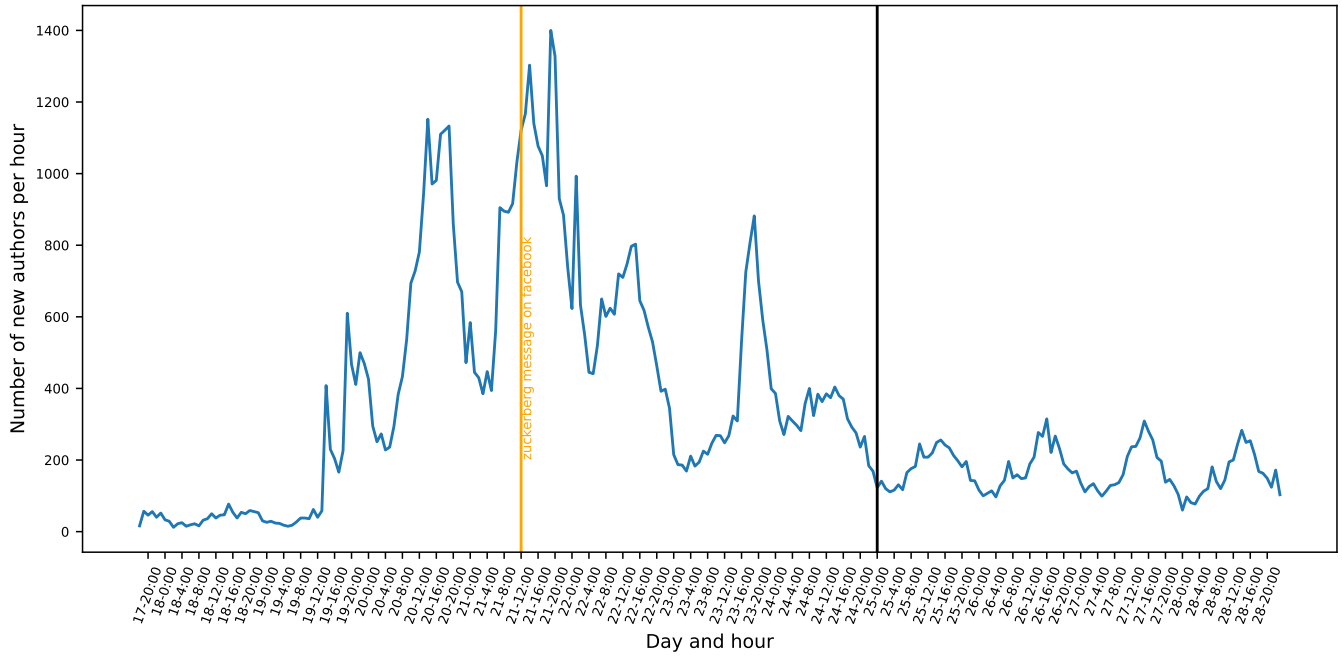

Figure 2.1: New authors time history

3 Network properties

3.1 Degree distribution

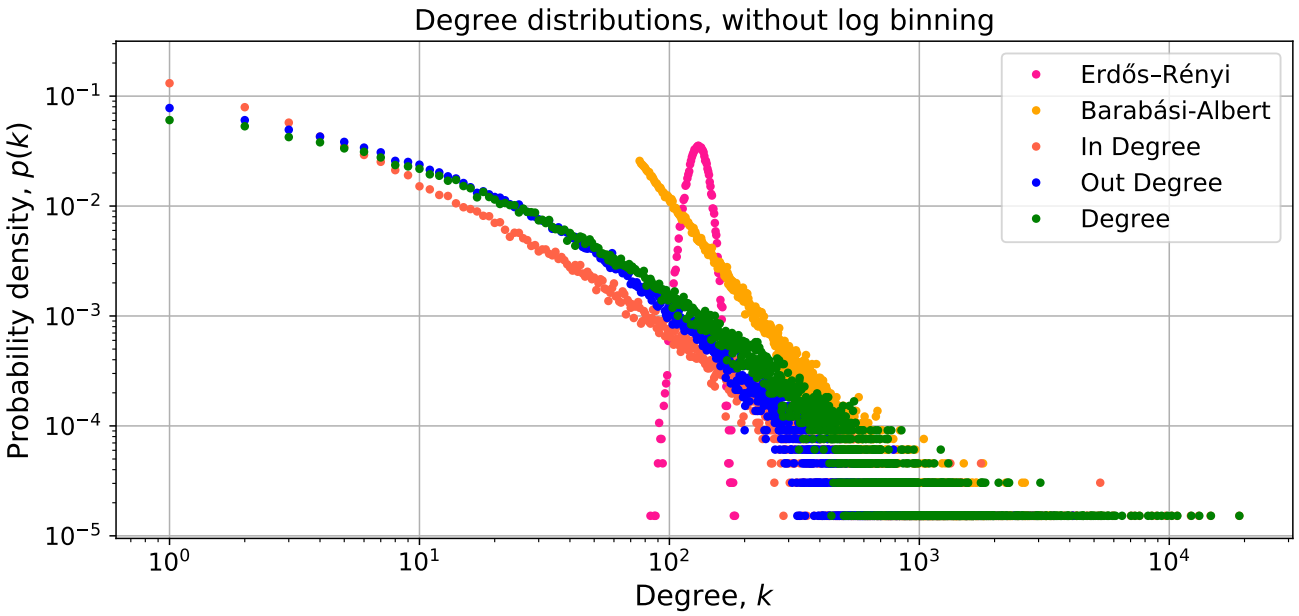

Figure 3.1: New authors time history

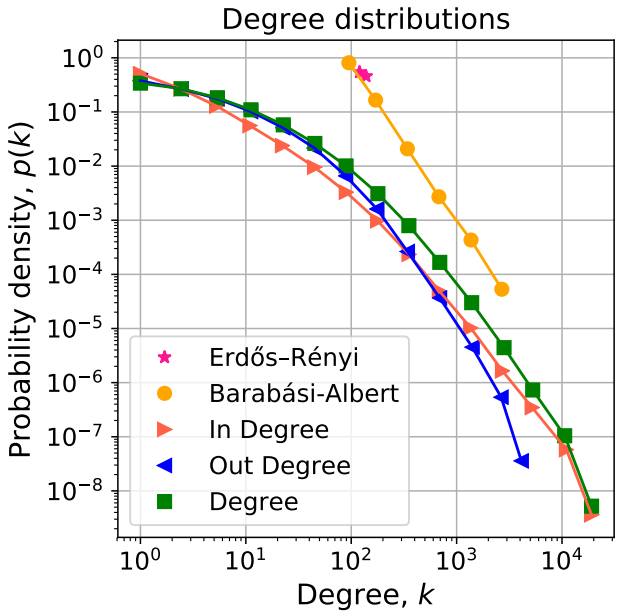

Figure 3.2

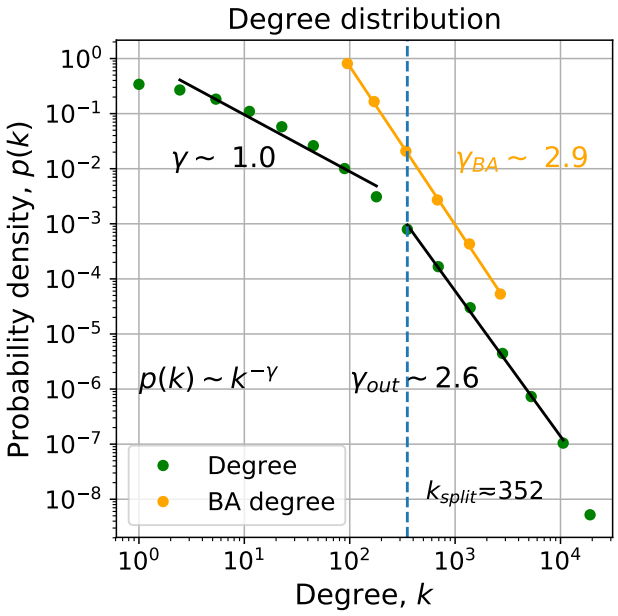

Figure 3.3

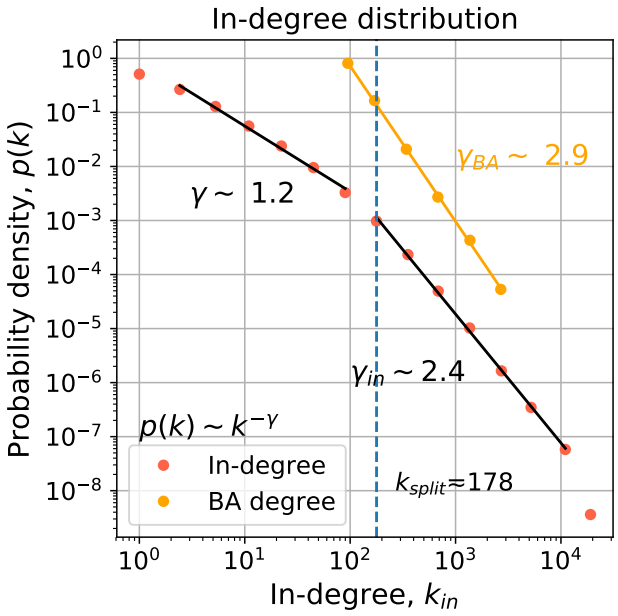

Figure 3.4

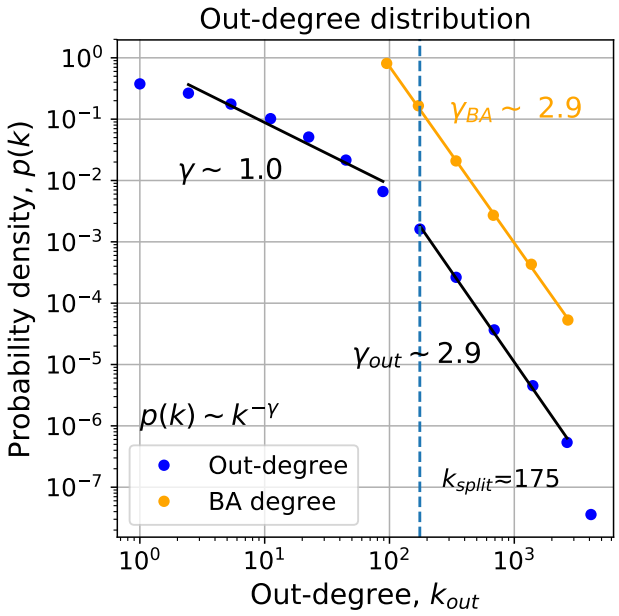

Figure 3.5

3.1.1 Random graphs

In order to generate an Erdos-Renyi random network we have choosen a "linking probability" p using the average degree of the original undirected network, by using eq. 3.1.

$$p_{ER} \approx \frac{\langle k \rangle}{N} = \frac{57}{65729} \approx 0.001 \tag{3.1}$$

Each new node of the random network generated with the Barabasi-Albert model has been attached to the other nodes with a number of links m equal to the average degree of the original network, considered undirected:

$$m = 2 \langle k \rangle = 76 \tag{3.2}$$

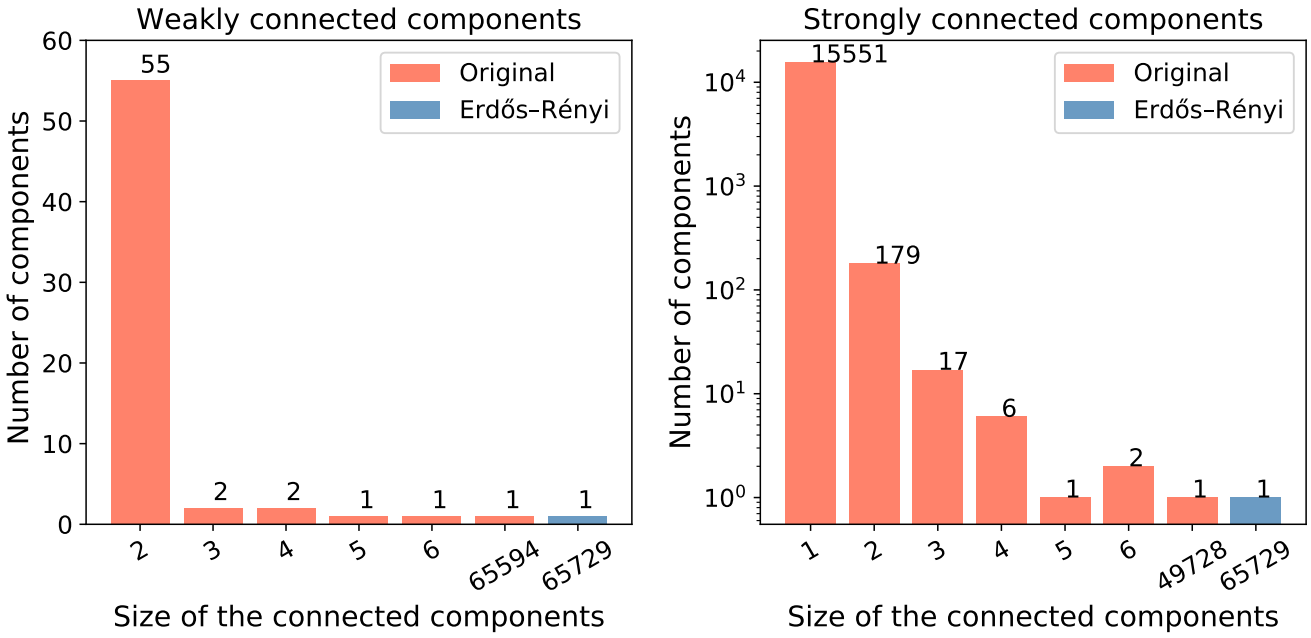

Figure 3.6: Connect components

3.2 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 [3]. 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:

$$\langle d \rangle = \frac{\sum D_i}{n} \ , \ \sigma(\langle d \rangle) = \frac{s}{\sqrt{n}} \tag{3.3}$$

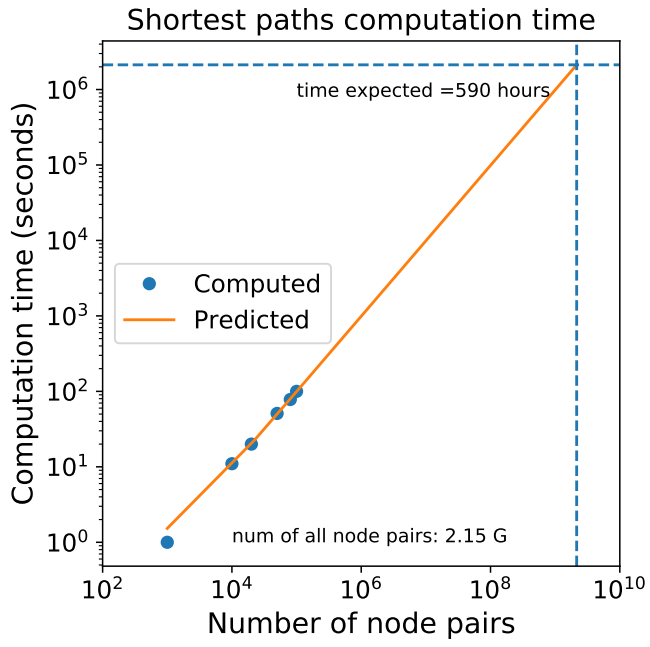

Figure 3.7: Shortest paths computation time by number of pairs

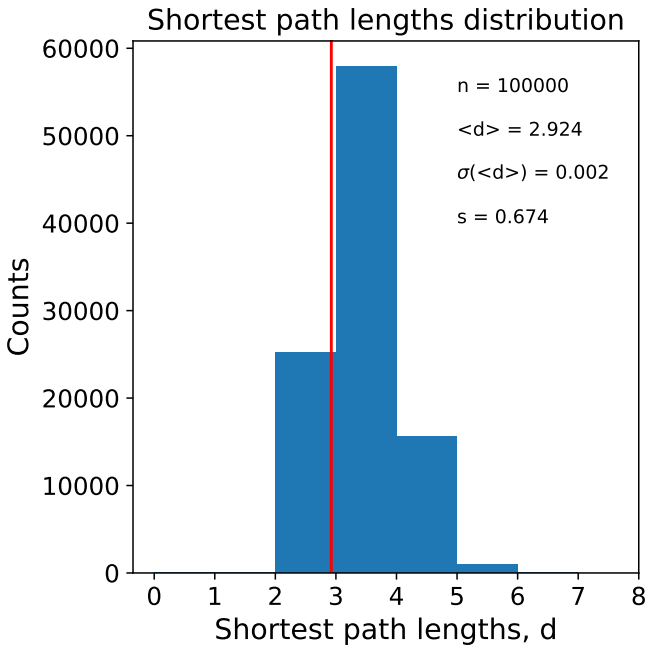

Figure 3.8: Shortest paths distribution

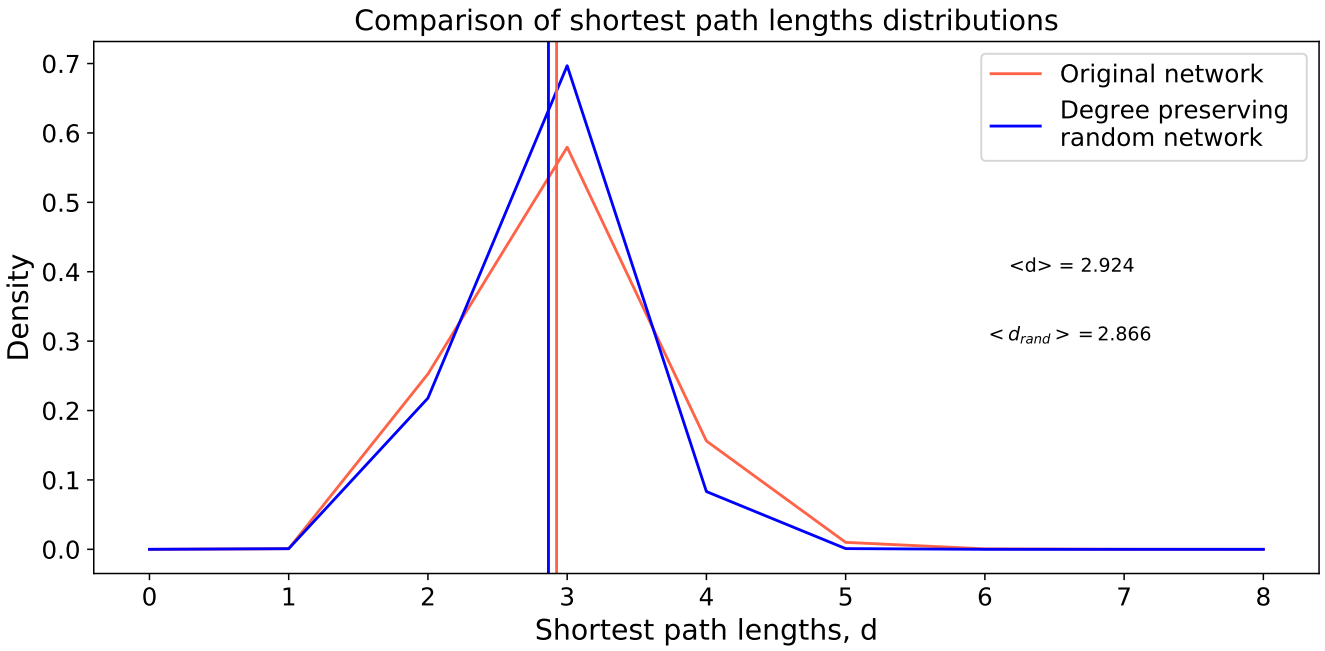

Figure 3.9: Shortest paths distributions comparison between the original largest connected component and a random network with degree preservation.

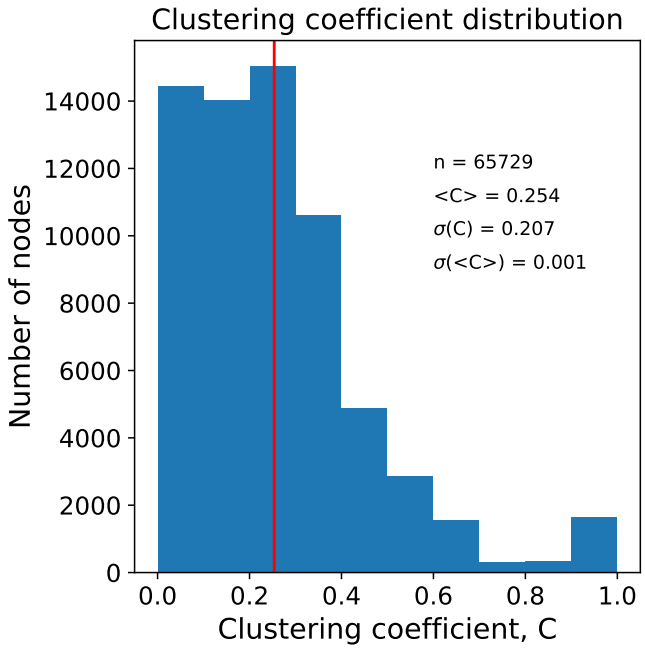

Figure 3.10: Clustering coefficient distribution

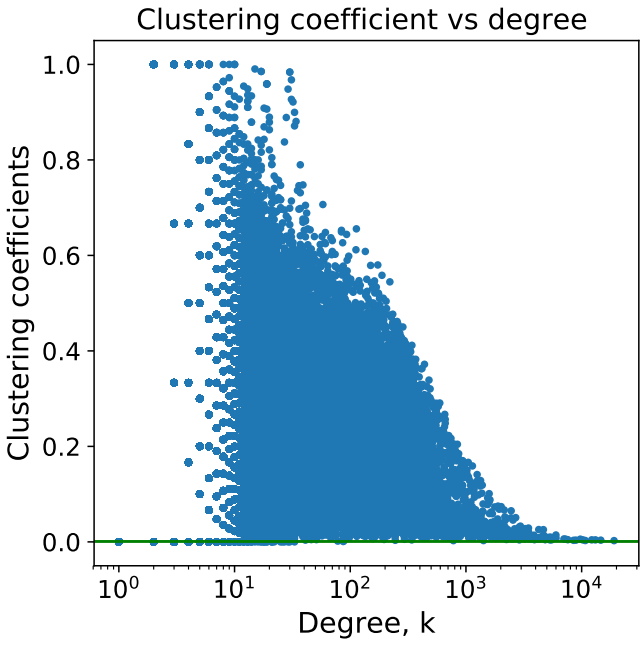

Figure 3.11: Clustering coefficients as function of the degree

3.3 Hubs analysis

The 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 the 30 biggest hubs are represented 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 an obvious 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:

$$\text{Interest} \equiv \frac{\text{in-degree}}{\#\text{followers}} \tag{3.4}$$

This measure represents the percentage of followers that being interested in the scandal had published a tweet about the subject.

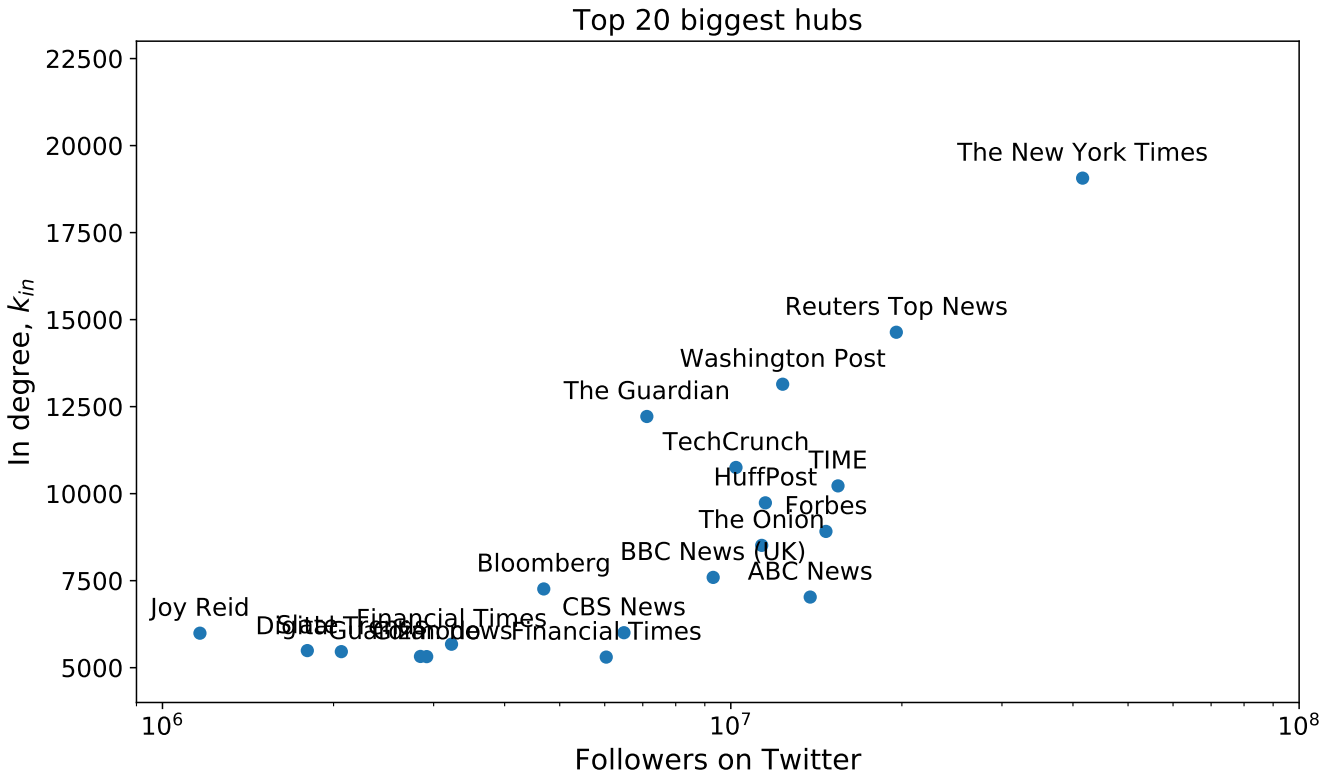

Figure 3.12

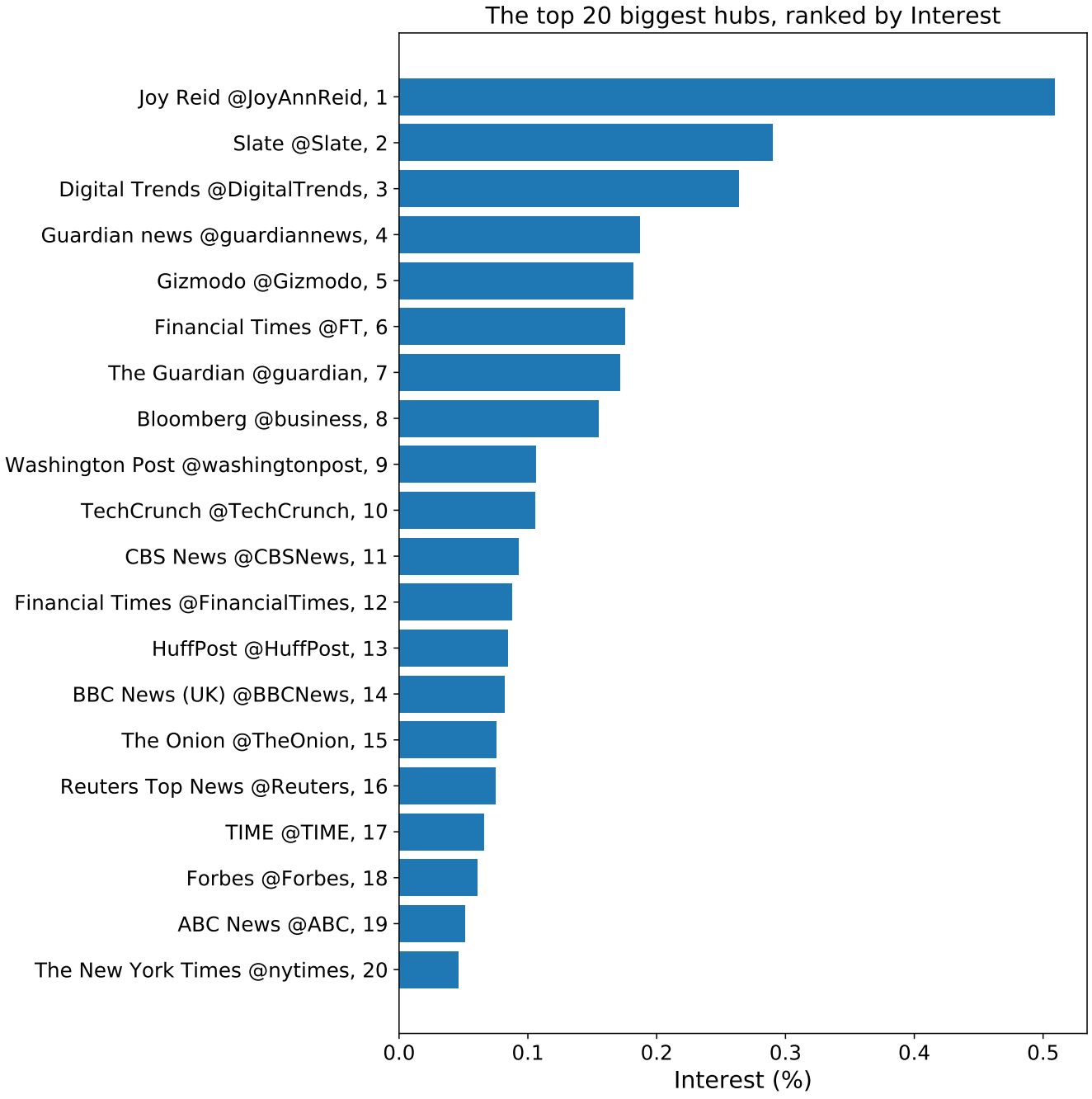
The top 20 biggest hubs, ranked by Interest
Joy Reid @JoyAnnReid, 1
Slate @Slate, 2
Digital Trends @DigitalTrends, 3
Guardian news @guardiannews, 4
Gizmodo @Gizmodo, 5
Financial Times @FT, 6
The Guardian @guardian, 7
Bloomberg @business, 8
Washington Post @washingtonpost, 9
TechCrunch @TechCrunch, 10
CBS News @CBSNews, 11
Financial Times @FinancialTimes, 12
HuffPost @HuffPost, 13
BBC News (UK) @BBCNews, 14
The Onion @TheOnion, 15
Reuters Top News @Reuters, 16
TIME @TIME, 17
Forbes @Forbes, 18
ABC News @ABC, 19
The New York Times @nytimes, 20
0.0
0.1
0.2
0.3
0.4
0.5
Interest (%)

Figure 3.13

4 | Network dynamics

5 | Communities discovery

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. For every model, a comparison between the three networks will be provided.

6.1 SI model

6.2 SIS model

6.3 SIR model

6.4 Threshold model

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.

[3] 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.