Network Robustness in the US Airports Infrastructure

CS224W - Project Report

Network Analytics

Paul Magon de La Villehuchet
Stanford University
paulmlv@stanford.edu

Abstract

Networks are present everywhere: in our relationships, in infrastructures, in technology and even in biology. Therefore, the analysis of networks will deeply improve our understanding of the world in which we live. Network theorists have developed several models to explain the properties of the networks surrounding us. The most renown model is the Random Graph model developed by Paul Erdos and Alfred Renyi. More recently, Duncan Watts and Steven Strogatz have developed a more realistic model known as Small-World. However, both this models, even though useful, still lack an essential property observed in many real networks: the degree distribution is not a nice bell curve but rather a power-law. Networks exhibiting this property are called scale-free networks and have been discovered recently (late 1990s). the objective of this project is to study the properties of scale-free network using an empirical graph: the US Airports network.

Introduction

The goal of the project is to study empirically the properties of scale-free networks using a real network based on the airport infrastructure in the US. Specifically, the project is a three steps process:

1. Characterize the nature of the US airports graph
2. Measure the robustness of the network
3. Derive efficient strategies to protect the network against attacks

In this report, we detail all the results and methods used for every step. Related work and references are presented in the last section

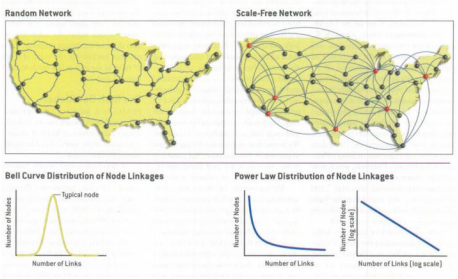

Figure 1: Scale-Free vs Random, extracted from [1]

1. Nature of the US airports graph

1.1. The data

1.1.1 Dataset motivation

In [1] article, A. Barabasi and R. Albert take the example of the airport infrastructure versus the highway infrastructure to explain the fundamental differences between random graphs and scale-free graphs. This prompted me to look for airport data to work on scale-free networks and the class website provided us with a good dataset so I decided to use it.

1.1.2 Dataset description

The data used in this project is a network of all US-Airports. Specifically, the network consists of a weighted edge list containing ids between two airports and the number of seats between the airports in 2002 (sum of the number of seats for all flights between the two airports). Therefore, using this data, the graph will be undirected and weighted.

1.1.3 Dataset collection

There was no data collection for this network as the dataset is available online. Indeed, our objective here is not to ap-

ply some algorithms to a new set of data but to learn about scale-free properties on an empirical network. Furthermore, I do like planes and airports therefore I think this data is great to explore properties of scale-free networks.

1.2. Scale-free networks

1.2.1 Characterization

Most real-life networks are dominated by very few nodes having almost unlimited number of connections, called hubs, while the vast majority of nodes have very few connections, hence the term scale-free. Such networks are extremely present in social networks and infrastructure networks. Mathematically, these networks are characterized by heavy-tail distribution unlike most generative models of networks: the tail is much bigger and the nodes at the end are statistically significant. In Random graphs and Small-World networks, the contribution of very node to the graph is relatively equivalent: the distribution of the degree usually follows a bell-curve decaying exponentially with the degree k. However, in the case of scale-free networks, the degree distribution follows a power-law: $\mathbb{P}(k) \propto k^{-\gamma}$. Empirically, the exponent value, γ, is between 2 and 3 and is a key characteristic of the network.

1.2.2 Generative mechanism

To understand why most real-life networks are scale-free, the authors in [2] developed a generative model , preferential attachment, leading to a scale-free network. The key idea of preferential attachment is that as the network grows, new nodes are more likely to connect to high-degree nodes. Specifically, the algorithm is in two steps:

> Growth: start with m_0 nodes and at each step t add 1 node with m edges.
>
> Preferential Attachment: connect edge j with node i with probability $\frac{k_i}{\sum k_k}$ with k_i the degree of node i.

Using this algorithm, the authors prove that the distribution of the resulting network is $\mathbb{P}(k) \propto k^{-3}$

1.3. The US Airports graph

1.3.1 General statistics

In this part, we considered the unweighted version of the graph, i.e the edges have all weights equal to 1. We will take into consideration weighting of the edges in the next part. The statistics we considered are: $|V|$ the number of nodes, $|E|$ the number of edges, ρ the density of the graph and $\bar{C}$ the average clustering coefficient.

Table 1: Summary statistics of the airport graph

Statistic	Value
$\|V\|$	500
$\|E\|$	2980
ρ	0.024
$\bar{C}$	0.62

As can be seen, this graph is characterized by an extremely high clustering coefficient for a rather low density. To explore the properties of this graph, I will compare our graph to "gold-standard" models seen in class: Erdos-Renyi, Watts-Strogatz. I will also explore the degree distribution of the graph.

1.3.2 Comparison with "gold-standards" models

Erdos-Renyi: The first model we can use to benchmark our network is the random graph generator from Erdos-Renyi. We obviously expect the graph to have different properties as the airport graph "should" be scale-free (more on that in the next subsection).

Table 2: Comparison between the airport graphs and the Erdos-Renyi model

Statistic	Airport	Erdos-Renyi
$\|V\|$	500	500
$\|E\|$	2980	2980
ρ	0.024	0.024
$\bar{C}$	0.62	0.023

As expected the properties are extremely different. We can see the big difference in the average clustering coefficient. This is not surprising as Erdos-Renyi graph are known to have low clustering compared to real-world data.

Watts-Stogratz: The second model we saw in class was the Small-World model. To generate a Small-World model graph, we used the method seen in assignment 1 that consists of:

1. Start with a ring of nodes
2. Connect each node to its neighbors's neighbors
3. Choose at random a non-connected pair of nodes and connect them

We saw in class that there the Watts-Strogatz model is characterized by a higher clustering coefficient than Erdos-Renyi therefore we expect the statistics to be closer to one another.

Table 3: Comparison between the airport graph and the Watts-Strogatz model

Statistic	Airport	Watts-Strogatz
$\|V\|$	500	500
$\|E\|$	2980	2980
ρ	0.024	0.024
$\bar{C}$	0.62	0.074

Even though the results are better for the Watts-Strogatz model, the model still does not capture the nature of the airport graph. Therefore, this leads me the believe that the nature of the airport graph will be indeed scale-free. To try to prove this statement, we will compare the graph with our third and last model: a scale-free model where the parameter is the exponent of the distribution.

Scale-Free: In the initial article, A. Barabasi and R. Albert mentioned that most of the networks encountered in the real world were scale-free with exponents ranging from 2 to 3. To try to understand the airport graph, we will compare its properties with networks generated using a power-law distribution. We decided to choose 3 values of the exponent γ: $1.5, 2, 3$.

Table 4: Comparison between the airport graph and the Scale-Free model

Statistic	Airport	$\gamma = 1.5$	$\gamma = 2$	$\gamma = 3$
$\|V\|$	500	500	500	500
$\|E\|$	2980	2875	1331	489
ρ	0.024	0.023	0.011	0.0039
$\bar{C}$	0.62	0.30	0.19	0.0027

The scale-free models give us results that are quite satisfying because all the numbers are in the right order of magnitude. Therefore, this leads us to believe that the nature of the airport graph is scale-free. Furthermore, from the experimentation presented above, we expect the graph to have a low exponent as $\gamma = 1.5$ got the closest results. In conclusion, the closest generative model seen in class to our airport graph is the scale-free model. To go deeper in our understanding of the airport graph, we will analyze its degree distribution.

1.4. Degree characterization

1.4.1 Degree distribution

The degree distribution is **the** key characteristic of scale-free networks. Figure 2 represents the degree distribution of the US Airports graph along the degree distribution of some generated Erdos-Renyi graph. As we can see, the nature of the distribution is completely different for both models and the intuition that the airport follows a power-distribution seems confirmed.

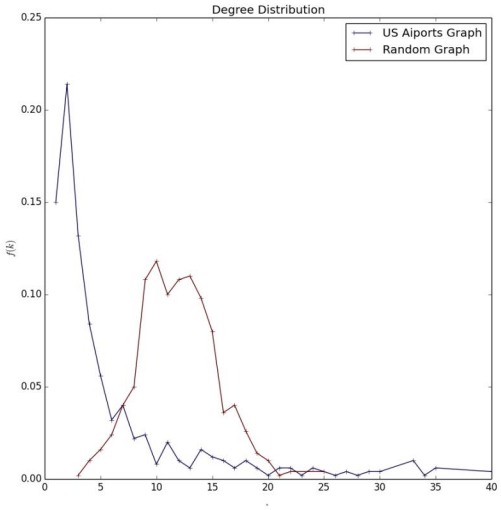

Figure 2: Degree distribution

1.4.2 Exponent estimation

To estimate the exponent of the graph, we used 3 methods seen in class. We applied these methods to both the weighted and the unweighted matrix. The intuition is that the power-law should be even more dominant in the weighted matrix as the degree of big airports will be inflated by the volume of traffic (number of seats).

Exponent fit: To determine the exponent of the power distribution, we fitted the degree-distribution in a log-log scale and then fitted a linear regression. The equation we fit is:

$$\log(\mathbb{P}(k)) = -\hat{\gamma}\log(k) + \hat{\beta}$$

This method is unreliable: the tail of the degree distribution is messy as can be seen been in Figure 3.

CCDF: The second method we used was to fit a linear regression on the Complimentary CDF. Specifically, we compute $\mathbb{P}(\deg > k)\ \forall k$ and then fit a linear regression on this distribution. The equation we fit is:

$$\log(\mathbb{P}(\deg > k)) = -\hat{\gamma'}\log(k) + \hat{\beta}$$

The value of the fitted coefficient $\hat{\gamma'}$ verifies $\hat{\gamma'} = 1 + \hat{\gamma}$. The interesting learning from this method is that even though the degree distribution is characterized by a heavy tail, a power law does not seem to fit perfectly as we can see that the CCDF is not a straight line in a log-log scale.

Maximum Likelihood Estimation: Finally, the last method we used was the maximum likelihood estimator.

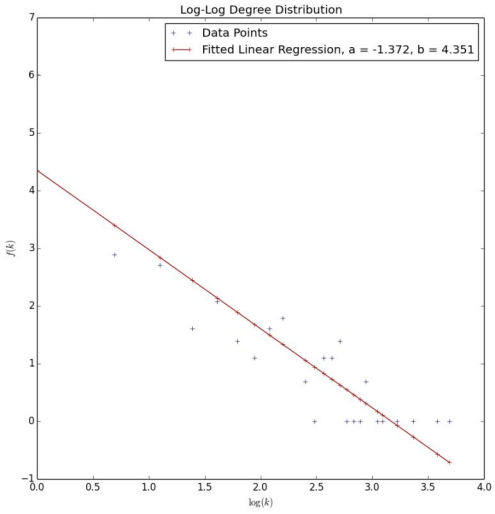

Figure 3: Fitted linear regression on the degree distribution

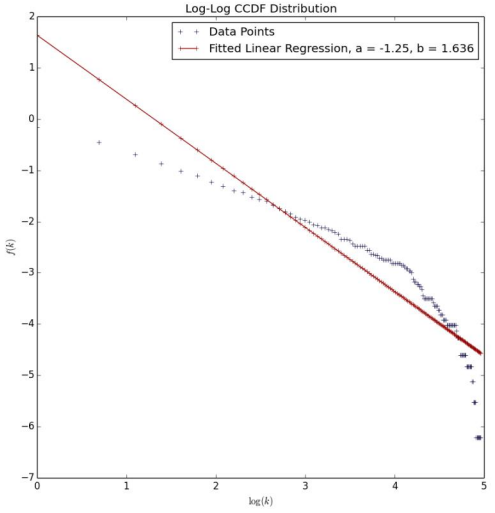

Figure 4: Fitted linear regression on the CCDF

Assuming the minimum value of the degree to be 1, the estimator is given by:

$$\gamma_{MLE} = 1 + \frac{n}{\sum_{i=1}^{n} \log(d_i)} \text{ with } d_i, \text{ degree of node } i$$

The following table summarizes the different estimations of the exponent.

Table 5: Exponent Estimation

Method	Weighted matrix	Unweighted matrix
Exponent fit	1.372	0.993
CCDF	2.25	1.96
MLE	1.64	1.15

1.5. Conclusion

We can draw 3 main learnings from this first part: first, the US Aiport graph is indeed characterized by a heavy-tail distribution. Second, we can see that this distribution does not seem to be completely scale-free as the plot of the CCDF shows some non-linearity in the log-log scale. Third, we can compute estimates of the exponent of the distribution. As was expected, those exponents are higher with the weighted matrix and are conform to empirical results from other scale-free networks (at least using the CCDF estimator, we find that γ is between 2 and 3). With this characterization done, we can move to the second part of the project where we will study the properties of robustness.

2. Network Robustness

2.1. General scheme

The goal of this part is to measure network robustness. Network robustness is an extremely important characteristic for infrastructure networks (and networks in general). Indeed, in big networks, nodes will eventually fail and thus impact the network as a whole. Knowing how robust and vulnerable is the network to such failures is key in protecting it. More precisely, in this part, we define two types of failures: first, random failures simulate random defects in the network. For instance, in the case of airports, that could be one airport being closed due to a an unexpected storm. Second, attacks simulate targeted attacks against the network with the objective of destroying the network. In the case of airports, that would be targeted attacks (by a malicious entity such as a rogue state) against specific airports to undermine the functioning of the US airport system as a whole. To measure robustness, I used the the relative size of the biggest connected component. The process to measure robustness was the following:

1. Choose some ordering of the nodes, relevant to the type of failure.
2. Eliminate nodes according to the ordering and measure the relative size of the biggest component at each iteration.

For random failures, the ordering I chose was simple: just pick nodes at random. For attacks, I decided to use centrality as an ordering of the nodes. The next part details which centrality measures were used. Finally, we repeated this procedure on the airport graph and on a generated Erdos-Renyi (used a benchmark).

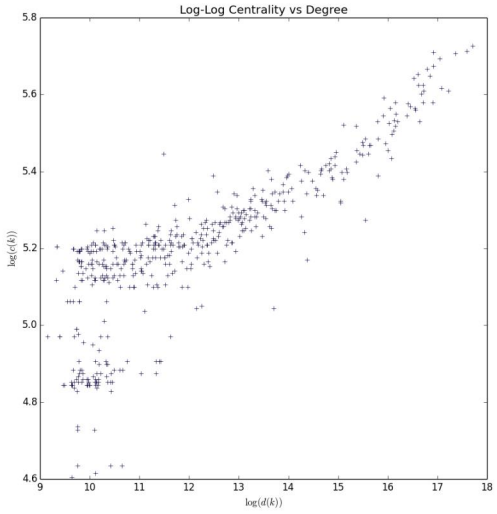

Figure 5: Correlation between harmonic centrality and degree

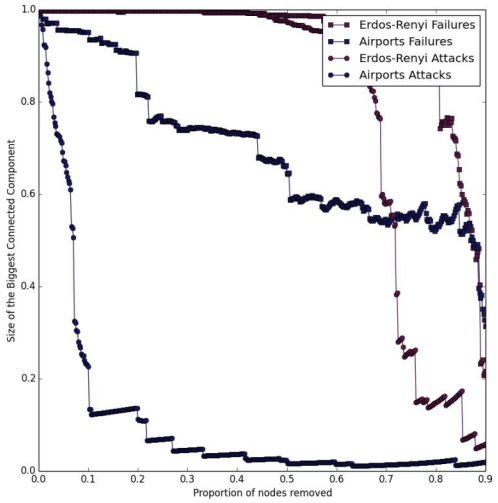

Figure 6: Robustness of the network

2.2. Measure of centrality

I used three measures of centrality, c, to determine which nodes to attack

1. Degree of the node: $c_i = \deg(i)$. This simple measure is extremely effective to compute and powerful in the case of scale-free networks.

2. Harmonic centrality: $c_i = \sum_{j \neq i} \frac{1}{d(i,j)}$ with $d(i,j)$ the shortest path between i and j. Harmonic centrality measures how much nodes need to rely on other nodes to convey information. Unsurprisingly, harmonic centrality is strongly correlated with the degree.

3. Eigenvector centrality: $c_i = \frac{1}{\lambda} \sum_{(j,i) \in E} c_j$ Eigenvector centrality measures the influence of a node in the network.

2.3. Results: Airports vs Random graphs

In this first subsection, we only compare the result of failures of the network to failures in our benchmark graph. Furthermore, to simulate attacks, we used the degree measure as this is the most obvious measure.
The results from this experiment are both conform to intuition on the behavior of each graph: the Erdos-Renyi is sensitive in a similar way to both attacks and failures with attacks being slightly more effective. On the other end, the airport graph is quite insensitive to failures while being extremely sensitive to attacks. These behaviors are not surprising and follow the results in [2]. However, I was expecting the airport graph to be more robust against failures than the Erdos-Renyi graph and the graph shows that in both case the Erdos-Renyi is strictly superior. A few elements can help us understand this discrepancy: first of all, the rather small number of nodes can be an explanation. Second, the robustness metric used is the size of the biggest connected component. Other metrics are also used to measure network robustness.

2.4. Results: Impact of Centrality on Robustness

In this subsection, we assess the impact of centrality on the robustness of the network. As we can see in Figure 7, the results are quite conform to intuition. First, all metrics measure how important and central is a node therefore the scale-free network is extremely sensitive to attacks. Furthermore because of the high correlation between the harmonic centrality and the degree (Figure 5), the results are extremely similar in both cases. Finally, the eigenvector centrality is slightly different than the other two as it emasures the number of paths of infinite length hence smoothing the sharpness of the degree distribution, it is closer to the random failures scheme.

2.5. Conclusion

In conclusion, centrality plays a key role in the robustness of the network: scale-free networks are particularly sensitive to targeted attacks against the network and centrality both provides the optimal strategy to attack the network and

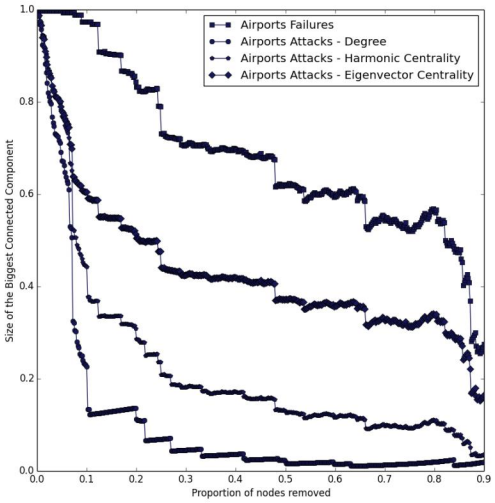

Figure 7: Robustness of the network

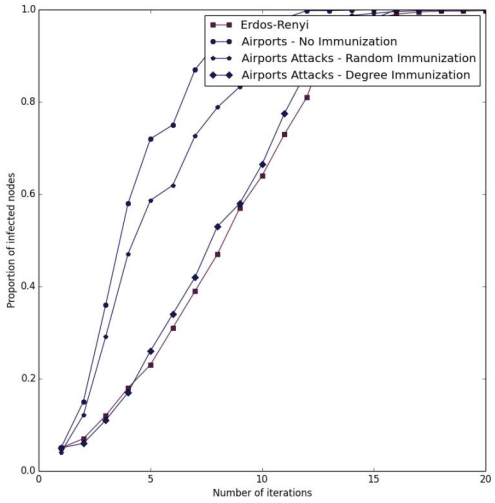

Figure 8: Propagation of the epidemy

to defend it. Furthermore, our results show that the simpler measure of centrality, i.e the degree if the network is the most effective. This is not really surprising since the robustness metric we used is the size if the biggest component therefore it is extremely sensitive to hubs: removing them will cause the network to be segmented hence explaining the results. To go further in our understanding of the network robustness and centrality, I decided to simulate the propagation of an epidemic in the network in the last part.

3. Efficient strategies against attacks

3.1. General scheme

In the last part, we determined that network centrality plays a key role in network robustness and is therefore the key to deriving efficient strategies to protect networks. To understand how these concepts are linked, I decided to simulate the propagation of an epidemic on the US Airports graph. The epidemic process was the following:

1. Start by infecting τ nodes
2. At each step, every node infects its neighbors with some probability q

We decided to test the model against three strategies:

1. No immunization
2. Immunize τ nodes at random
3. Immunize τ nodes based on the degree

3.2. Results

In the above simulation, we chose $\tau = 5\%$. Moreover, we did not count immunized nodes when computing the proportion of infected nodes. The results of this experiment are conform to our intuition. First of all, the scale-free graph converges to full epidemic much faster than a random graph used as a benchmark. Furthermore, as with failures, targeting the highest degree nodes is the most effective strategy to protect the network since it reduces considerably the time to reach full contagion.

3.3. Conclusion

In conclusion, our most efficient strategy will simply be to protect the most connected nodes. One element that can be added is that in the case for which the graph is too big to compute the entire degree distribution (such as the Internet), one efficient way to determine the most connected nodes is simply to use random walks starting at random as these random walks will most likely end-up on highly connected nodes.

4. Conclusion

The goal of the project was to study the properties of the scale-free networks using an empirical dataset, the US Aiports dataset. I first studied the property of this dataset and found that it has indeed the properties of scale-free graphs even though its distribution is not perfectly a power-distribution. In a second part, I studied the robustness of the network using various measures of centrality to determine efficient ways to attack the network. Using these results,

I derived efficient strategies to protect the network against attacks. Even though I was not able to to do everything I hoped for, I enjoyed working in this project and enjoy the class in general.

5. Related work - Contributions

Being alone, I worked on everything myself.

Related work:

- A. Barabasi, and E. Bonabeau. *Scale-Free Networks*. Scientific American (2003)
  In this paper, the authors develop a new model of networks, scale-free networks. Most real-life networks are dominated by very few nodes having almost unlimited number of connections, called hubs, while the vast majority of nodes have very few connections, hence the term scale-free.

- R. Albert, and A. Barabasi. *Statistical Mechanics of complex networks*. Review of Modern Physics (2002)
  In this comprehensive paper, the authors address all the shortcomings of the first paper and lay-out all the mathematical theory underlying scale-free graphs. It was extremely helpful to understand the theory underlying scale-free networks.

- R. Albert, and A. Barabasi. *Error and attack tolerance of complex networks*. Review of Modern Nature (2000)
  The authors of this paper develop some tools to measure the robustness of the scale-free networks to failures and attacks. Specifically, robustness was measured for scale-free networks and benchmarked against Random Graphs. This paper was extremely helpful for the second part of the project when studying properties of centrality and robustness.

- CS224W Lectures notes
  I used a lot the lecture notes in my project, especially for centrality and scale-free characterization. It would be unfair not to mention it.

6. References

[1] A. Barabasi, and E. Bonabeau. *Scale-Free Networks*. Scientific American (2003)

[2] R. Albert, and A. Barabasi. *Statistical Mechanics of complex networks*. Review of Modern Physics (2002)

[3] R. Albert, and A. Barabasi. *Error and attack tolerance of complex networks*. Review of Modern Nature (2000)