Network Analysis
Wikipedia Graphs

Federico Fontana, Lorenzo La Corte

May 24, 2023

Chapter 1

Introduction

This group of datasets includes three wikipedia page-page networks based on three different topics: chameleons, crocodiles and squirrels.

Nodes represent articles from the English Wikipedia in December 2018, edges reflect mutual links between them.

1.1 Our Datasets

Since the original dataset didn't contain any information about the topic or the title of the specific articles, we created a **tool** that lets us generate new graphs starting from a wikipedia link, scraping pages up to a certain depth (in the case of Directed Crocodile the depth was set to 2) and saves the information about the title of each page.

The tool performs a BFS on wikipedia pages, and it only considers links contained in the `div` with id `bodyContent`. Furthermore, in an effort to try to keep a common topic of the articles forming the graph, we decided to add a set of keywords that the content must contain to be considered a relevant page (and thus be scraped) (in this case, the only keyword was "crocodile")

It is worth noting that these checks are only performed on pages that are being scraped, and not on the destinations of the `a` tags. Thus, the pages pointed by the last "layer" of pages may not have a `div#bodyContent` or may not even contain any of the keywords provided to the script.

1.2 General Overview

These are some basic facts in this dataset:

	Chameleon	Squirrel	Crocodile	Directed Crocodile
Nodes	2277	5201	11631	49314
Edges	31421	198493	170918	167272
Density	0.012	0.015	0.003	0.0001

Chapter 2

Centrality

In this analysis we will focus on three main measures:

1. **Degree Centrality:** local measure, is the total number of neighbours at distance one.
2. **Betweenness Centrality:** global measure, is the number of geodesic paths passing through a node.
3. **Closeness Centrality:** globabl measure, is the harmonic mean distance from a vertex to the other vertices.

In this section we will mainly be interested in the ranking of the nodes for these metrics. In the case of directed crocodile, we will also be taking a look at the titles of the highest ranked pages with respect to these metrics.

2.1 Chameleon

These are the results regarding centrality in this dataset:

Node	Degree	Node	Betweenness	Node	Closeness
1976	0.3216	1939	0.3587	1939	0.4677
1939	0.2974	1976	0.1593	1976	0.4357
1741	0.2873	1741	0.0903	1741	0.4135
2263	0.2333	2249	0.0819	1862	0.3905
2246	0.1766	1911	0.0697	2249	0.3889
1356	0.1414	2246	0.0667	2246	0.3830
220	0.1353	1708	0.0554	2263	0.3804
2249	0.1199	1846	0.0516	2164	0.3755
1714	0.1195	1862	0.0493	1356	0.3693
1333	0.1181	1923	0.0353	652	0.3688

Most Central Nodes: We can notice that 1741, 1976 and 1939 dominate the top 3 of all metrics. These are probably the most influential nodes in the network.

Bridges: The nodes 1911, 1708, 1846 are only ranked for their betweenness. Although the may not be as central as the first three nodes, these are crucial for the control flow of the information.

2.2 Squirrel

These are the results regarding centrality in this dataset:

Node	Degree	Node	Betweenness	Node	Closeness
4346	0.3663	4346	0.1345	4346	0.5202
5112	0.3563	5112	0.0905	5112	0.5145
4365	0.2736	4903	0.0405	4903	0.4753
4903	0.2540	4303	0.0394	4365	0.4726
4303	0.2405	4841	0.0270	4303	0.4720
5095	0.2315	4989	0.0255	5095	0.4580
4419	0.2209	3290	0.0247	4544	0.4577
5033	0.2171	5194	0.0243	5164	0.4569
5063	0.2107	5164	0.0234	4864	0.4548
4322	0.2078	4365	0.0223	4419	0.4547

Most Central Nodes: We can notice that 4303, 4903, 5112 and 4346 dominate the top 5 for all metrics. In respect to Chameleon, we can observe that degree is higher for the top ten of nodes: all of them are linked with at least 20% of the network. This suggests that this graph has more giant hubs.

Bridges: The nodes 4841, 4989, 3290 and 5194 are only ranked for their betweenness. As mentioned before, these are vital for the flow of information in the network and so for its failure resistence.

There is also a node (4365) with low betweenness (10th) but high closeness and degree. Among all the nodes reported in the table, this is the one node with the characteristics that best fit those of a **Free Loader**. This means that this node is central and very near to other central nodes, but without controlling the flow as much as other nodes in the table.

2.3 Crocodile

These are the results regarding centrality in this dataset:

Node	Degree	Node	Betweenness	Node	Closeness
11535	0.3049	11535	0.1608	11535	0.4876
9632	0.2368	11256	0.1436	9632	0.4663
10437	0.2235	11216	0.1010	11216	0.4551
10118	0.2104	10118	0.0819	10118	0.4429
11068	0.2004	9632	0.0771	10928	0.4404
7230	0.2000	10928	0.0606	8715	0.4317
11618	0.1830	11127	0.0554	11509	0.4315
11596	0.1821	10437	0.0524	11256	0.4277
10252	0.1712	10588	0.0419	11339	0.4267
11256	0.1685	10318	0.0379	10437	0.4247

The Center: We can notice that the node labeled with 11535 has the top rank in all of the three categories. This is the most important node in the network.

The Bridge: We can also notice that the node identified with 11256 has quite a central role: it is traversed by lot of shortest paths between nodes and is close to some of the other hubs. The interesting fact is, although it is top 2 ranked both in betweenness and closeness, it's only 10th by degree.

The Free Loader: the node 8715 has high closeness but is not top ranked for betweenness and degree.

2.4 Directed Crocodile

Due to the dimension of this graph, only the node degree centrality is reported. Luckily we have the possibility to associate nodes with their pages, which lets conduct a more interesting semantic analysis.

Node	Degree	Page Title
92	0.0428	Marine Biology
754	0.0337	Veganism
748	0.0321	Vegetarianism
66	0.0315	Bird
239	0.0308	Steve Irwin
65	0.0284	Fish
709	0.02777	Kebab
744	0.0253	Ethics of Eating Meat
107	0.0242	Venezuela
33	0.0242	Africa

The first thing we can notice is that the top 10 nodes in this ranking have a degree centrality that is roughly ten times lower than the the nodes with the highest degree in the other datasets. This is probably due to the fact that this

graph is comprised of many more nodes. In proportion, this means that hubs are smaller, as we will see in the next chapter.

From the semantic point of view there are a lot of results that we would expect:

1. **Steve Irwin:** "The Crocodile Hunter",
2. **Africa**, which has the largest population of Nile crocodiles.
3. **Venezuala**, where there are 5 species of crocodiles

Then there are less expected pages, but still related to...

1. **animals:** Marine Biology, Bird, Fish
2. **eating animals:** Veganism, Vegetarianism, Ethics of Eating animals.

And then there is **Kebab.**

2.5 Directed Crocodile Pruned

Unfortunately, the original dataset was too big to analyze its betweenness and closeness.

In order to rank this dataset with respect to the other centrality metrics, the dataet has been pruned using core decomposition. In particular all nodes with degree less than three are erased from the graph, and these are the results:

Node	Degree	Node	Betweenness	Node	Closeness
Vegetarianism	0.169	Vegetarianism	0.055	Crocodile	0.530
Veganism	0.167	Crocodile	0.053	Crocodylidae	0.530
List of Pork Dishes	0.154	Crocodylidae	0.053	Doi Identifier	0.512
Ethics of Eating Meat	0.153	Veganism	0.052	ISBN Identifier	0.509
List of Smoked Foods	0.150	Ethics of Eating Meat	0.044	PMID Identifier	0.507
Carnism	0.148	Carnism	0.041	PMC Identifier	0.504
List of Sausage Dishes	0.126	Fish	0.031	S2CID Identifier	0.503
Crocodilia	0.122	Crocodilia	0.027	ISSN Identifier	0.500
Dog Meat	0.122	Snake	0.027	Snake	0.499
List of Steak Dishes	0.121	List of Smoked Foods	0.026	Turtle	0.496

Due to pruning, the 10 nodes with the highest degree centrality show higher values with respect to the original dataset.

From the semantic point of view we can notice some patterns in the rankings:

1. **Degree Centrality** top 10 contains a lot of **lists**. This makes sense semantically because these pages contain a lot of links to other pages.
2. **Betweenness Centrality** top 10 contains nodes more related to the starting point (crocodile) from which the dataset was built.

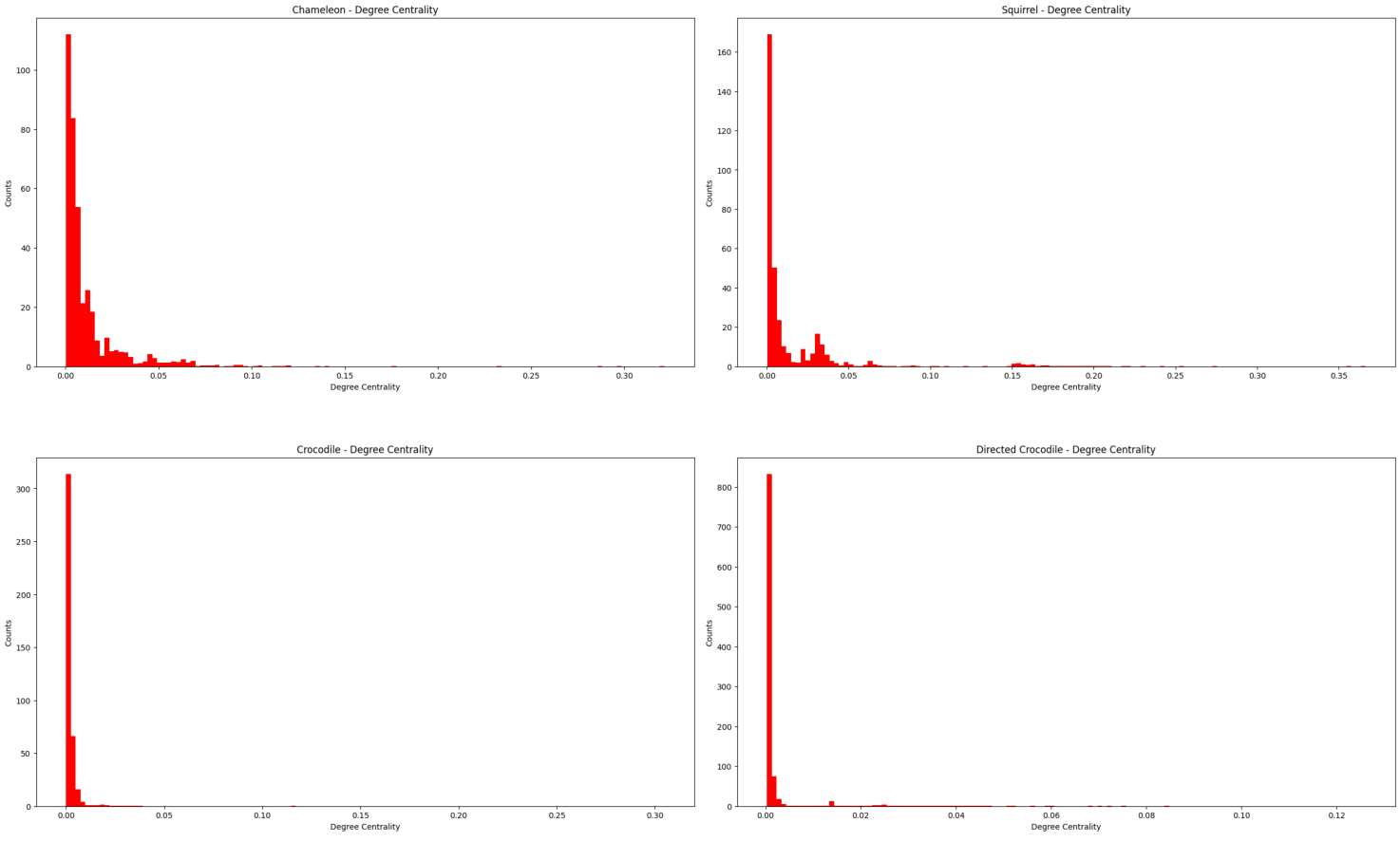
3. **Closeness Centrality** top 10 have some on-topic pages and a lot of **identifiers** linked by pages in the references section

In a sense, betweenness is the metric that most capture central nodes on a semantic point of view. Degree and Closeness centrality tables, instead, are well-related to their definitions and capture two kinds of pages in wikipedia which are meant to have those roles.

2.6 Degree Centrality Comparison

In general, there are few nodes with high degree and a lot of nodes with low degree.

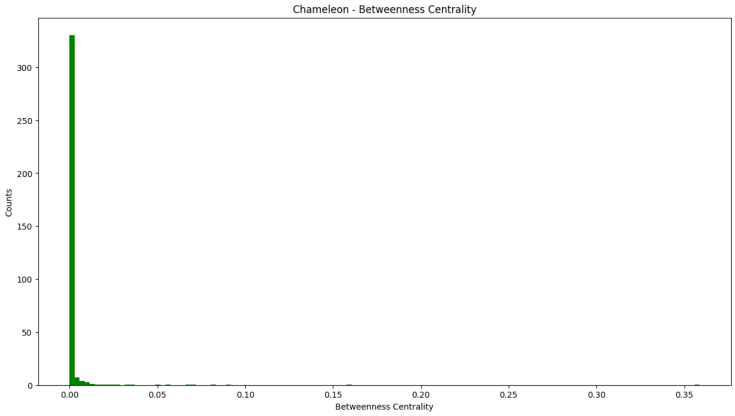
In this plot we can notice that as N increases, the degree distribution becomes more and more peaked in the lower values of the degree. This will be an aspect of paramount importance in the next chapter.

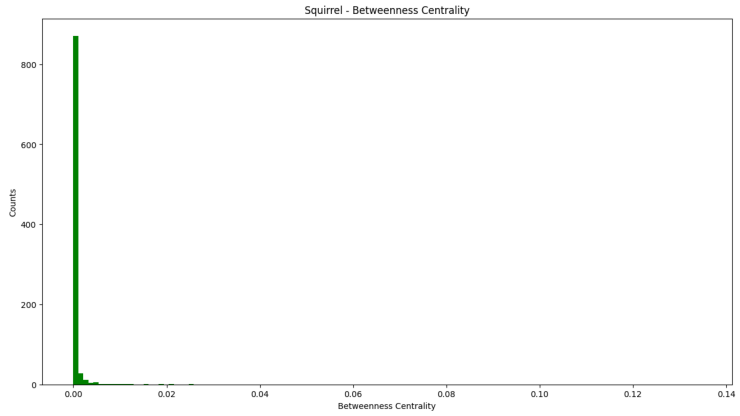


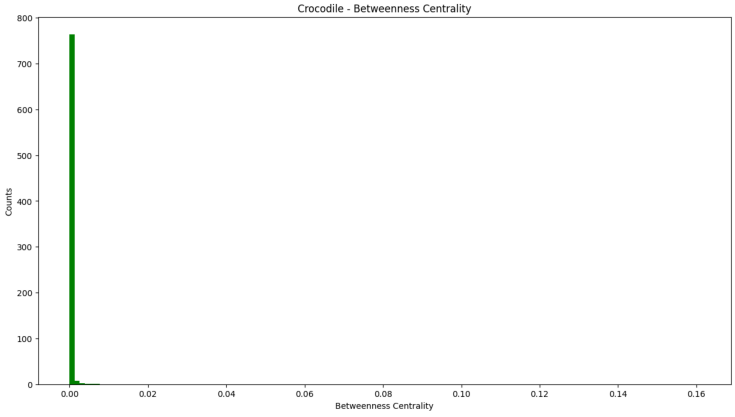
2.7 Betweenness Centrality Comparison

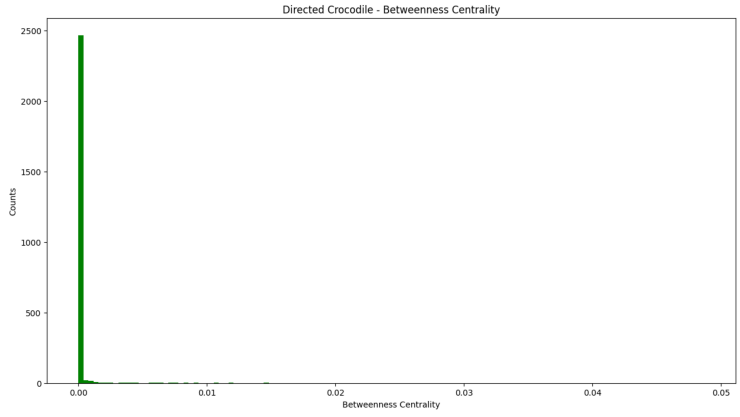
Betweenness shows the presence of:

1. lots of nodes with very low betweenness,
2. very rare nodes that play a crucial role in the structure of the graph.







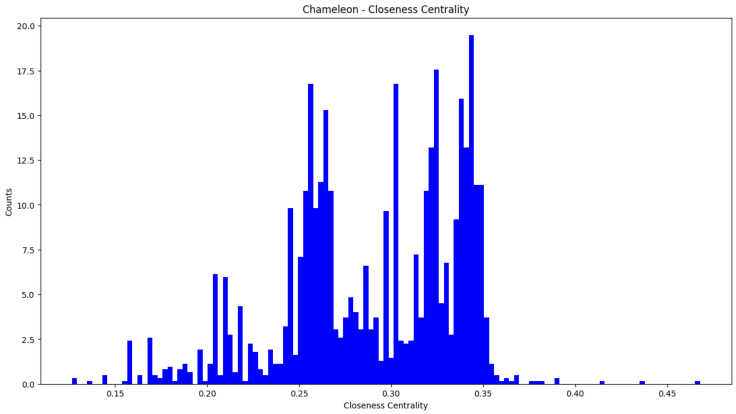


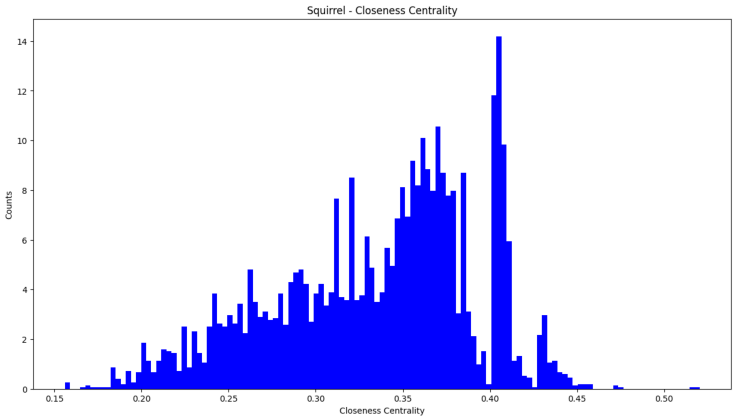
2.8 Closeness Centrality Comparison

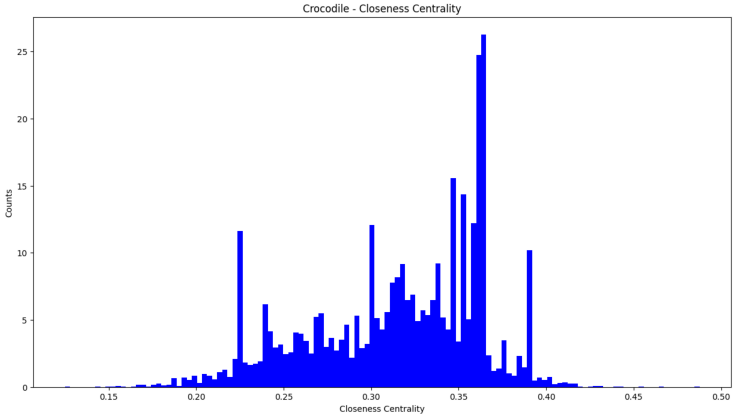
Closeness plots are interesting, especially for the differencies between:

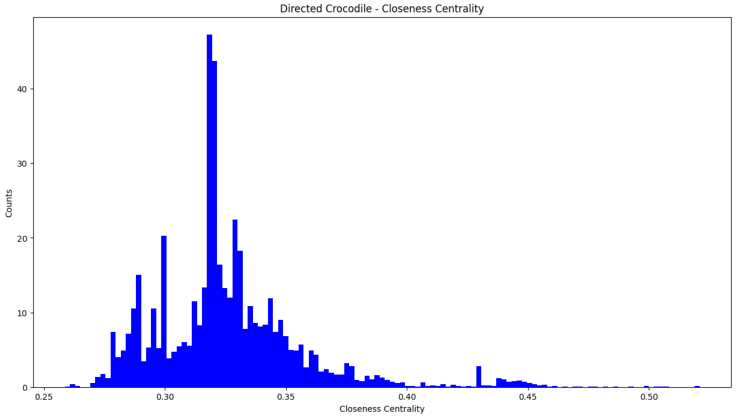
1. original datasets, which show distributions skewed towards higher values for this metric
2. our crocodile dataset, which shows less nodes with high closeness value. It is howvere interesting to notice that those nodes show higher closeness.

It's clear that in both the four datasets there are lot of nodes linked to hubs. For squirrel and our crocodile, the horizontal scale is wider. For squirrel is also clear that there are more nodes with highness centrality. So, we can define this dataset as the one with more central nodes in this sense.









Chapter 3

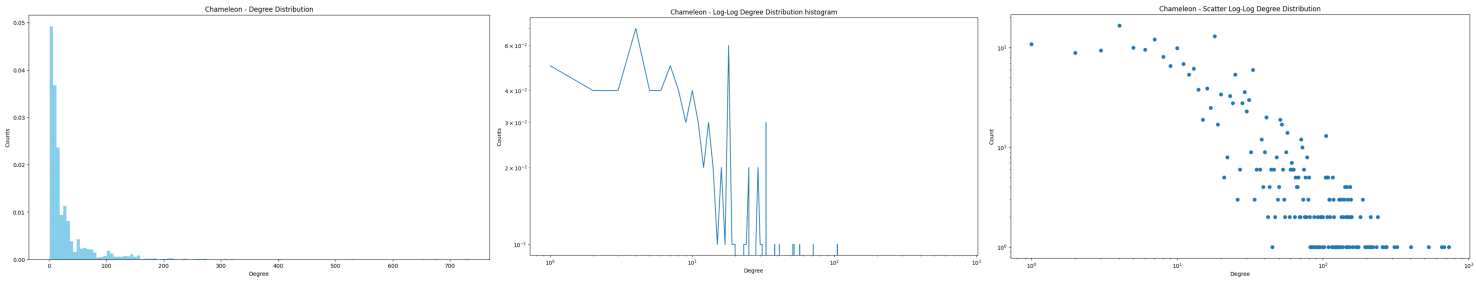
Degree Distribution

Analyzing the degree distribution and other properties of the network can provide insights into the underlying structure and dynamics of the system, and help us understand how it functions and evolves over time.

In many real-world networks, including the three we will analyze, degree distribution follow a power law, meaning that there are a few nodes with very high degrees (hubs) and many nodes with lower degrees. We already have some insights of the structure of these graphs from the previous degree centrality plots, but we will further analyze this aspect and look for the dataset with the clearest power-law characteristics.

3.1 Chameleon

This is the smallest graph, as it only has 2277 nodes and 31421 links. Despite its size, this graph provides valuable insights into the structural characteristics of this kind of graphs, that we expect to have similar structures



It's difficult to define the trend of the degree distribution.

If we look at the figure on the left, it appears as a power law with some irregularities.

In the log-log curve of the distribution the line tends to go down to low to zero

as the degree increases. When visualized on this scale, the degree distribution of the Chameleon graph appears to approximate a straight line.

It doesn't seem linear because it presents a lot of irregularities but, as a power law distribution, it has:

- few nodes with very high degree (hubs)
- many nodes with low degree

which is consistent with the ultra-small world regime.

By calculating the angular coefficient, denoted as θ, it is possible to determine the equation of the power law distribution. This finding further supports the notion that the Chameleon graph exhibits a scale-free structure, as we will try to prove.

3.1.1 Ultra-Small World Evidencies

If we analyze metrics, this graph shows lots of characteristics of a ultra-small world regime:

1. The degree distribution of the Chameleon graph shows that the majority of nodes have a degree between 4 and 10.

   There are also 109 nodes with degree 1 and few nodes with a higher degree, with the maximum degree being 732.

   Note that k_{max} is two (almost three) orders of magnitude bigger than k_{min}. This heterogeneity in node degrees suggests a presence of hubs, as we will discuss on further analysis.

2. The mean degree $\langle k \rangle$ is 27.59, and the variance is 2156.31.

   Such a high value for the variance suggests a lack of scale for this graph.

   As a matter of fact, if $N \to \infty$, we have that $\sigma \to \infty$ and $k_{max} \to \infty$.

   All moments, except for the mean (first moment), will tend to diverge, as a proof of the ultra-small world regime.

   We can also notice that $\langle k \rangle = 27.59 > \ln(N) = 7.73 > 1$, and so we can conclude that this graph is in the connected phase of the phase transition process.

3. Let's now analyze the mean distance among nodes. From a theoretical point of view, if $N = 2277$, the mean distance among nodes should be:

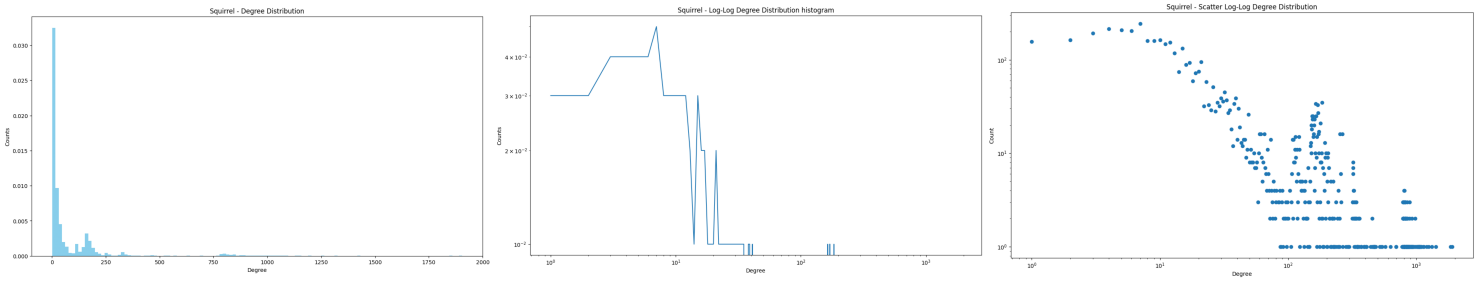
   (a) $\ln(N) = 7.73$, in a small-world regime,

   (b) $\ln(\ln(N)) = 2.05$, in an ultra-small world regime

   The empirical mean distance result is 3.56, which is closest to the expected value of this metric for the ultra-small world regime.

3.2 Squirrel

This graph has 5201 nodes and 198493 links and connected.



This graph shows a similar trend as Chameleon.

3.2.1 Ultra-Small World Evidencies

These are the evidencies we gathered by analyzing the graph's characteristics:

1. The most common degree is around 7, with 243 vertices having a degree of 7. The distribution has a long tail to the right, with degrees ranging up to over 1900.
   Furthermore, in the same manner as before, we can see that k_{max} is 3 orders of magnitude greater than k_{min}

2. There are 20 nodes with more than 1000 edges, which indicates the presence of a few hubs in the network.

3. The mean degree of the graph is 76.32, which is 1/25 of the maximum degree.

   Moreover, variance of the degree distribution is very large (26074.12); the fact that the variance is so high is a sign of the ultra-small regime, where this metric approaches infinity as N and so it doesn't carry any meaningful information.

4. From a theoretical point of view, if $N = 5201$, the mean distance among nodes should be:

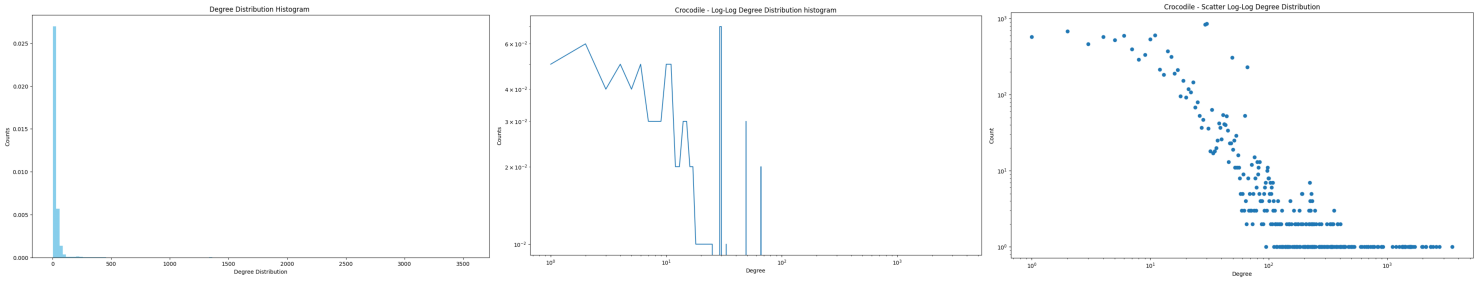
   (a) $\ln(N) = 8.56$, in a small-world regime,

   (b) $\ln(\ln(N)) = 2.14$, in an ultra-small world regime

   The empirical mean distance result is 3.09 and is thus closest to the ultra-small world regime. It is also worth noticing that this graph presents the smallest gap between theoretical an empirical values of the mean distance.

3.3 Crocodile

This is the largest graph among the three, it has 11631 nodes and 170918 links, which means that it is a relatively large graph.



As we can see in the image above, apart from some points, the degree distribution is becoming more and more an approximated linear model as the graph grows.

3.3.1 Ultra-Small World Evidencies

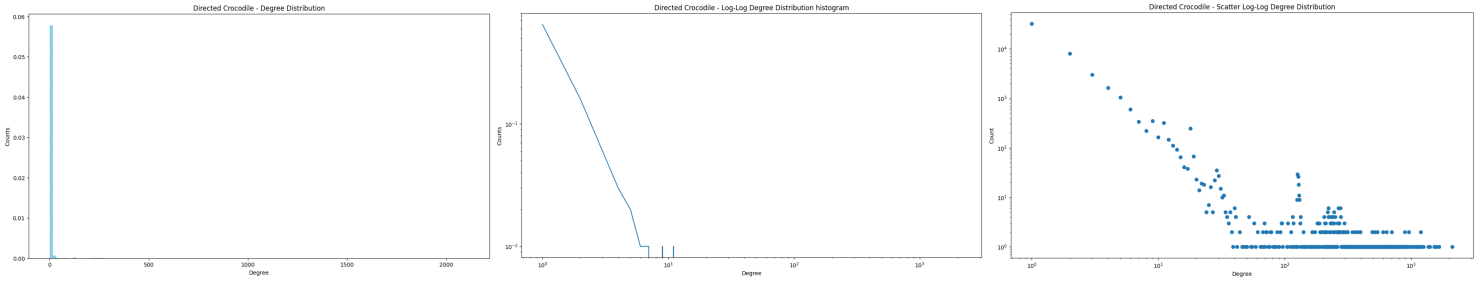
The analysis brought to these results:

1. k_{min} is 1, while k_{max} is 3546, three orders of magnitude bigger.
2. There are 34 nodes with more than 1000 edges, which indicates the presence of hubs in the network.
3. The mean degree of the graph is 29.39, which is relatively low compared to the maximum degree, indicating the presence of many low-degree nodes in the network.
4. The variance of the degree distribution is very large (11505), which indicates that the degree distribution is highly skewed and may not be well-described by a simple mathematical model. The fact that the variance is so high is also a sign of the ultra-small regime, where this metric tends to infinity and as such it does not carry any meaningful information.
5. From a theoretical point of view, if $N = 11631$, the mean distance among nodes should be:
   (a) $\ln(N) = 9.36$, in a small-world regime,
   (b) $\ln(\ln(N)) = 2.23$, in an ultra-small world regime

   The empirical mean distance result is 3.25, which is closer to the ultra-small world regime expected value.

3.4 Directed Crocodile

The custom-built dataset shows crucial differencies.



The degree distribution shown in the plot above above is clearly a power law. This kind of degree distribution is the clearest among the four datasets probably due to the size of the graph (almost 50000 nodes).

3.4.1 Ultra-Small World Evidencies

This graph shows characteristics of a ultra-small world regime:

1. There are 32031 nodes with degree 1 (k_{min}), while the highest degree (k_{max}) is 2112 and is 3 orders of magnitude bigger than k_{min}. Furthermore, many other nodes are connected with 1000 or more other nodes.
2. The mean degree $\langle k \rangle$ is 6.13, and the variance is 2271.90. This is thus another example in which nth moments with n higher or equal than 2 lose their meaning.

3.5 Take-Away Points

Considering the curves' trend as the size of the graph increases, the degree distributions plotted with logarithmic axes can be approximated by a straight line. This is a proof that these kinds of graphs are scale-free.

Although, for the smaller datasets, the degree distribution is not clearly the expected one. The equation of the power law is defined for $N \to \infty$, and so, as expected, datasets with more nodes better approximate a line in the log-log plot.

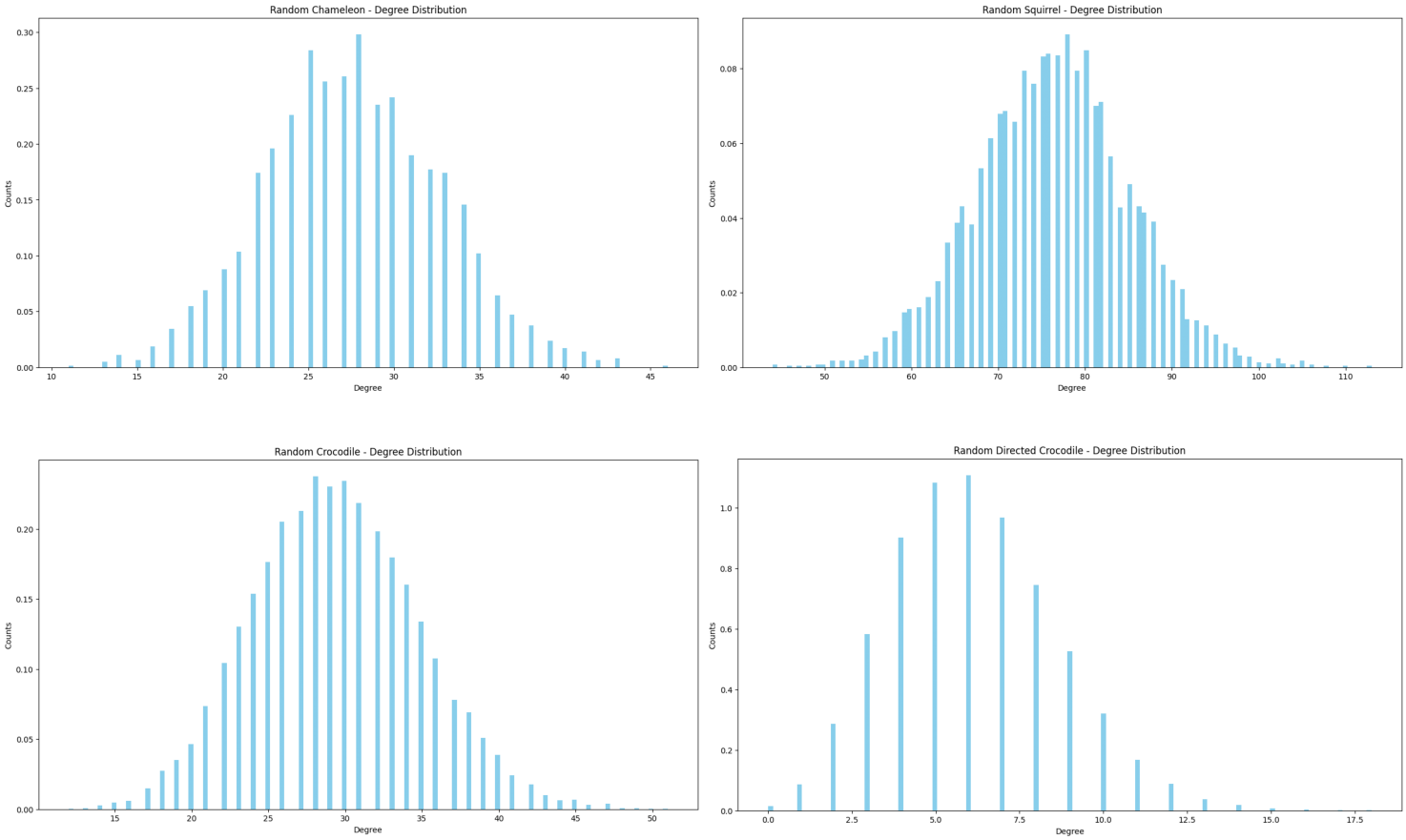
We confirm these hypotesys by analyzing the metrics:

1. All the four datasets show presence of hubs, shrunk distances and ultra-small world regime.
2. Surprisingly we can see that Crocodile, the largest original dataset, is not the one with metrics that best approximate the ones expected for the ultra-small world regime.
3. Directed Crocodile is the dataset with the clearest evidencies.

3.6 Degree Distribution of Randomized Datasets

A further experiment consist of:

1. building randomized versions of our datasets with same number of nodes and edges,
2. analyzing their degree distribution.



Results and clear: as expected, the degree distributions of the randomized versions of these graphs follow a bell shape.

Chapter 4

Transitivity and Clustering

This chapter is focused on transitivity and clustering metrics. These metrics are important in understanding the structure of a network and can reveal patterns that may not be immediately apparent.

The local clustering of each node in G is the fraction of triangles that actually exist over all possible triangles in its neighborhood. The average clustering coefficient of a graph G is the mean of local clusterings.

Graph transitivity is the fraction of all possible triangles present in G. Possible triangles are identified by the number of triads (two edges with a shared vertex).

The average clustering coefficient is derived from local measures, and thus it is easier to compute, but limited in use. On the other hand, transitivity is a global measure that is intrisecally harder to compute, but it usually carries more informative, especially when used to compare the metrics of graph models with real world graphs.

These are the values of these metrics for our datasets:

	Chameleon	Squirrel	Crocodile	Directed Crocodile
Triangles	1.029×10^6	2.879×10^7	1.87×10^6	7.218×10^6
Avg Clustering Coef.	0.481	0.422	0.336	0.235
Transitivity	0.314	0.348	0.026	0.127

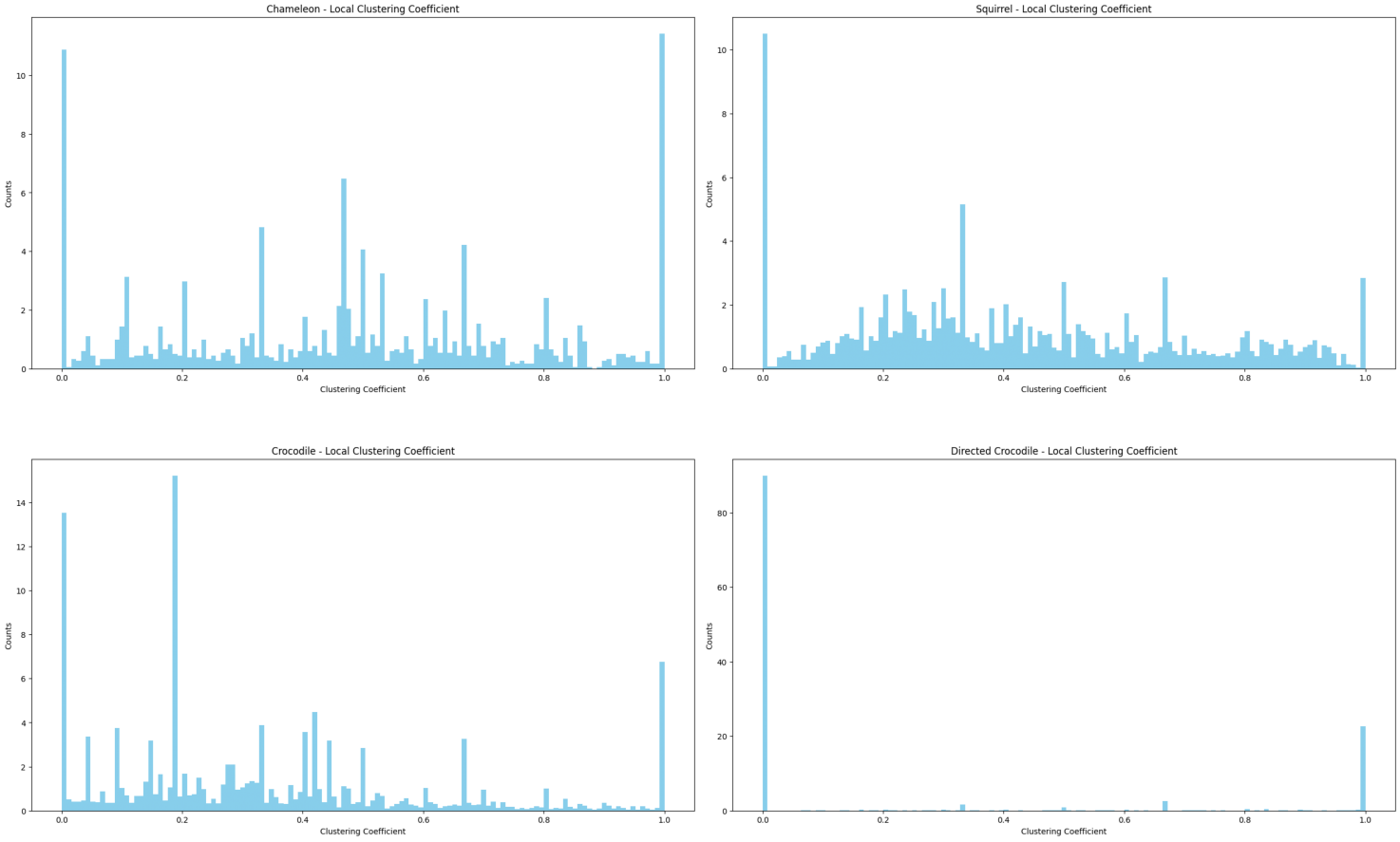
The number of triangles is obviously an absolute metric that also depends on how many nodes are present in the graph, so we are not so interested in this metric. It is however interesting to notice that Squirrel, which has half the nodes of Crocodile, has 10 times the triangles Crocodile has.

Concerning transitivity and clustering, we can notice that Crocodile is the dataset with the lowest rank, both in the Directed and Undirected cases. Regarding the other two, there's an interesting fact:

1. Chameleon is the dataset with the highest local clustering coefficient,
2. Squirrel is the dataset wihth highest transitivity.

The comparison between the average clustering coefficient and transitivity can provide relevant insights: Chameleon has a relatively high average clustering coefficient but a lower transitivity, suggesting that while nodes in Chameleon form local clusters, these clusters are not strongly interconnected on a global scale. Squirrel, on the other hand, has a lower average clustering coefficient but a higher transitivity, indicating a more globally interconnected network with fewer local clusters. So, Chameleon shows a higher tendency for articles to form local clusters, while Squirrel exhibits a higher level of global interconnectedness.

The same concept also applies for Crocodile and our version of this dataset, which shows a lower clustering coefficient but a higher transitivity.



Chapter 5

Assortativity

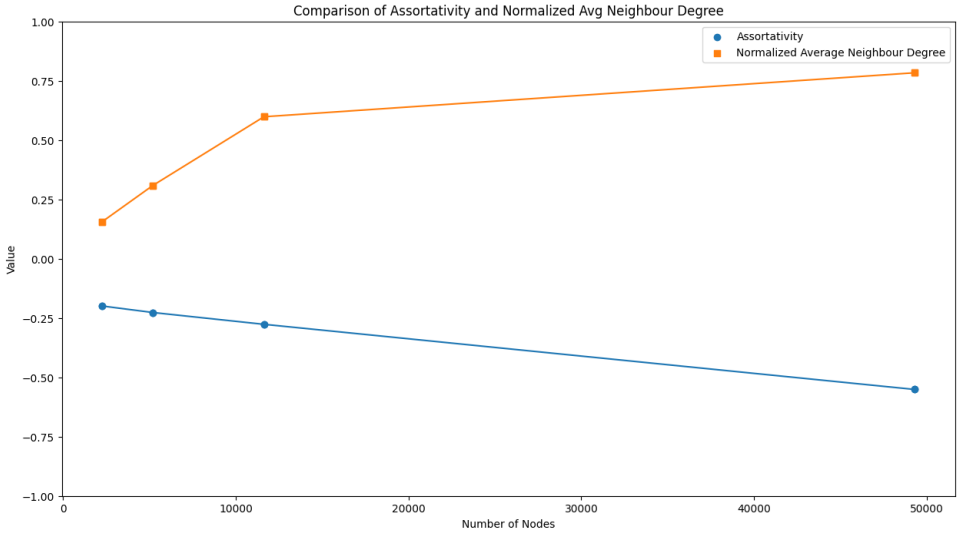
Assortativity in a network refers to the tendency of nodes to connect with other similar nodes, with respect to a property. The property we are going to analyze is the degree, and so assortativity will be:

1. positive, if nodes tend to connect with other nodes with similar degree,
2. negative, if they tend to connect with others with different degree.

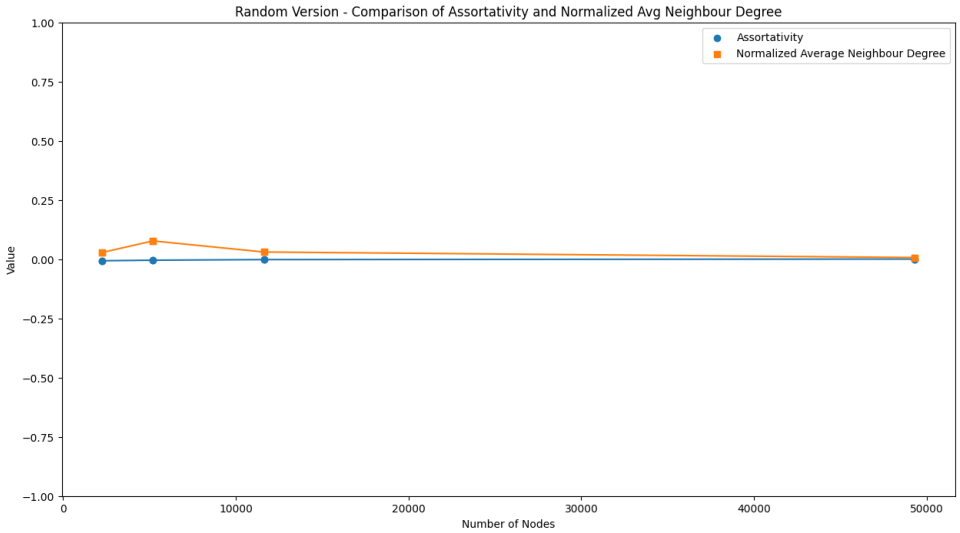
	Chameleon	Squirrel	Crocodile	Custom Crocodile
Assortativity	-0.199	-0.226	-0.276	-0.551
Average Neighbour Degree	156.492	308.767	598.264	783.607

All datasets show disassortativity: nodes tend to connect to dissimilar nodes over similar nodes. This suggests a **Hub-and-Spoke** network, which intuitively makes sense when thinking about the way these datasets were constructed.

Here we can have a look at the trend for the evolution of both metrics with respect to the nodes in the datasets:



It seems like a a linear trend for the assortativity, while the average neighbour degree follows a straight line until a certain amount of nodes, where it flattens out. In order to show the differences with random graphs, this is the correspondent plot on the randomized version of all the graphs:



As expected both of the metrics show values really close to 0 when computed on these graphs.

Chapter 6

Communities

In the following table we can see the partition of nodes which maximises the modularity (using Louvian heuristics) for each dataset. The table also shows the most important details for the best partition.

	Chameleon	Crocodile	Squirrel	Custom Crocodile
Number of communities	15	19	7	23
Modularity	0.691	0.689	0.403	0.654
Coverage of largest community	32,2%	20,8%	21,9%	16,2%

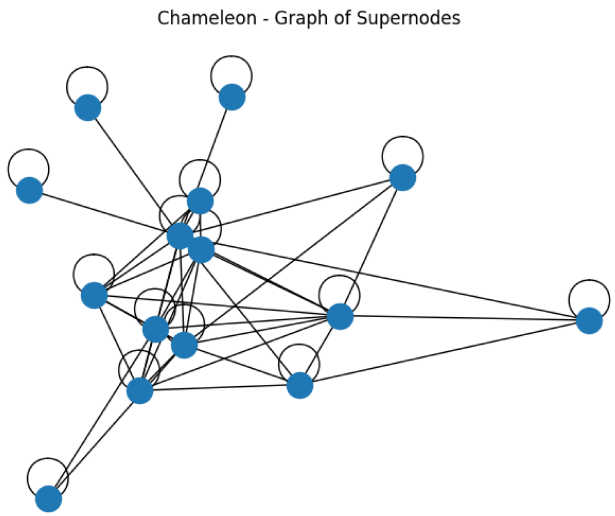
We can once again notice how Squirrel is particular: it has a very low number of communities:

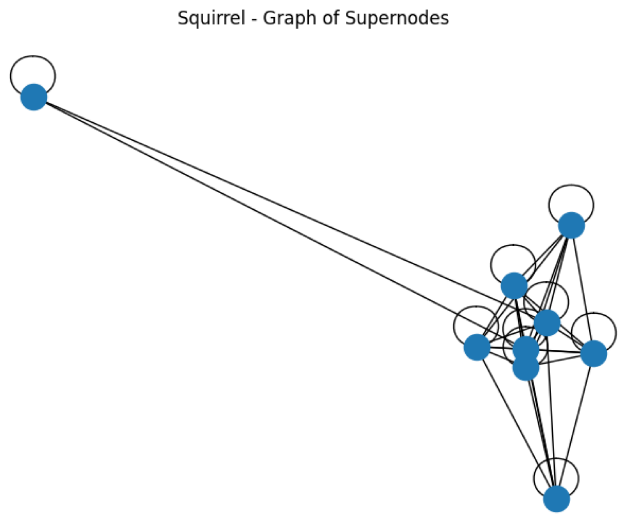
- although it has double the nodes of Chameleon, it has half the communities.
- communities don't differ much in sizes.

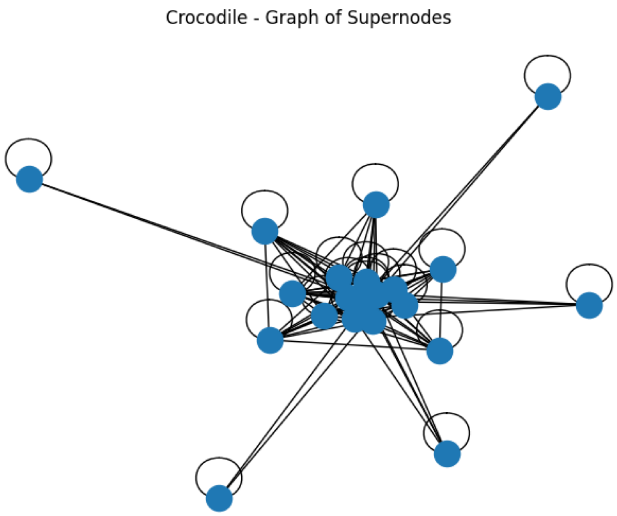
It is also interesting to notice that Squirrel has the lowest modularity (0.4) and that it is a lot lower than all of the others, which are in the range between 0.65 and 0.70. Modularity evaluates communities with respect to a random baseline, and this gives a quality measure to the partition. For this dataset, it seems that Louvain method hasn't found a better set of communities to enhance the value for this metric.

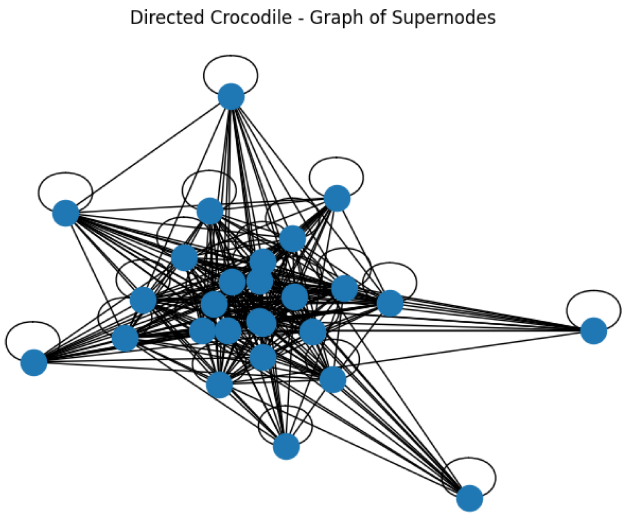
We built Louvain Supernodes Graphs for each dataset, which show the supernodes found by the Louvain Partition Method. The nodes belonging to each community have also been color coded and reported in colored-by-community graphs.

6.1 Louvain Supernodes Graphs

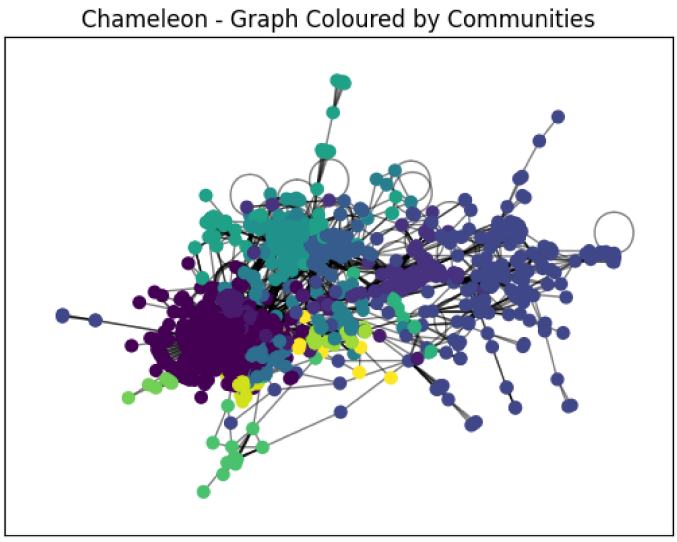


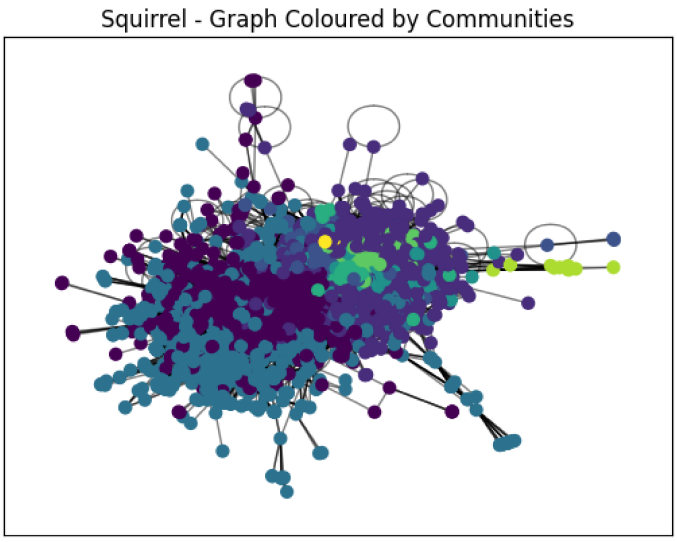


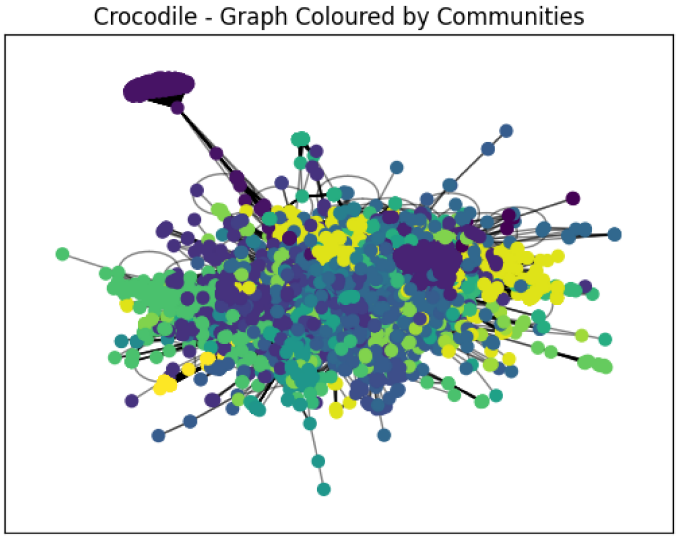


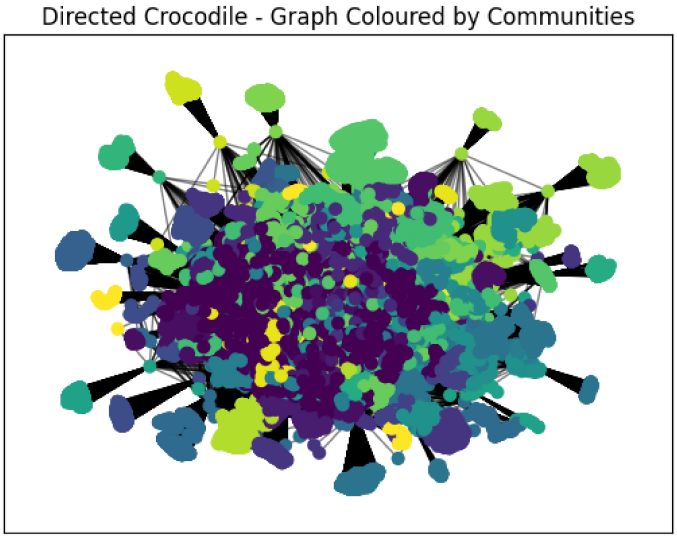


6.2 Graphs Coloured by Communities







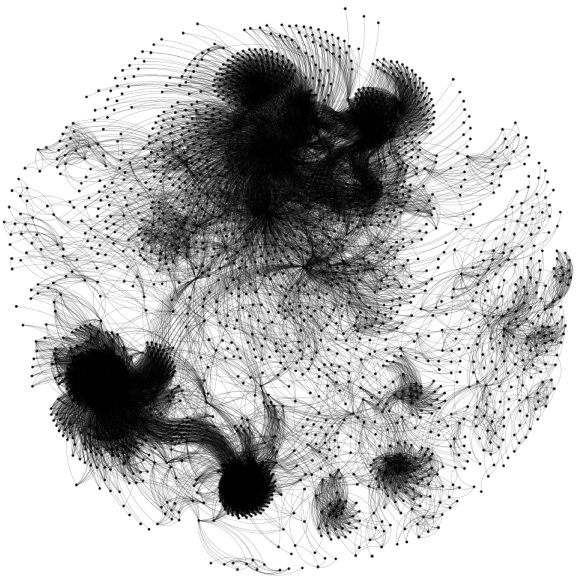


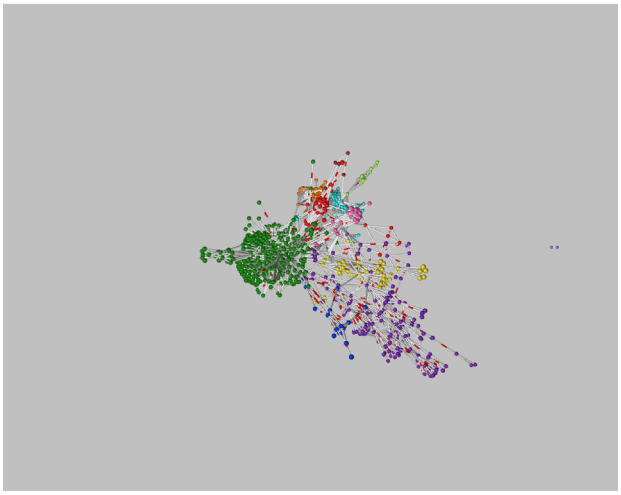
Chapter 7

Visualization

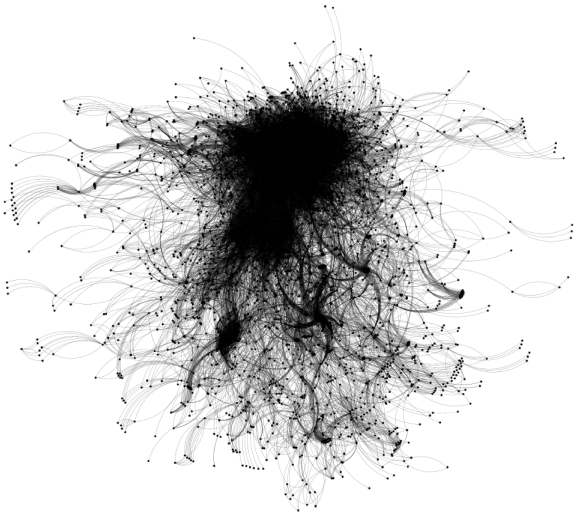
Although is not useful for the analysis, is interesting to conclude our report visualizing our graphs. The two visualizations are made with **Gephi** and **Graphia**.

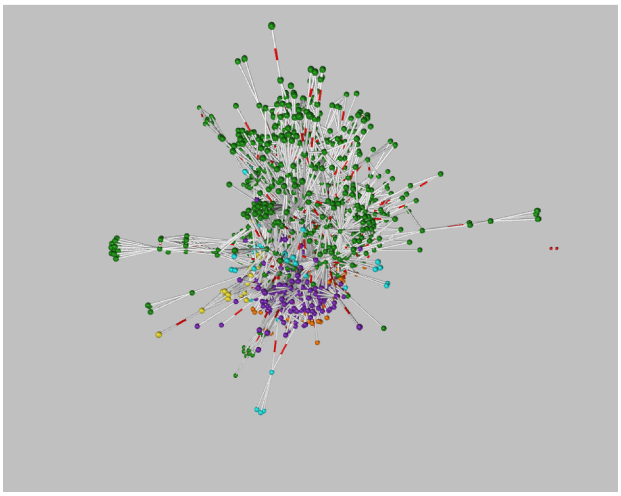
7.1 Chameleon



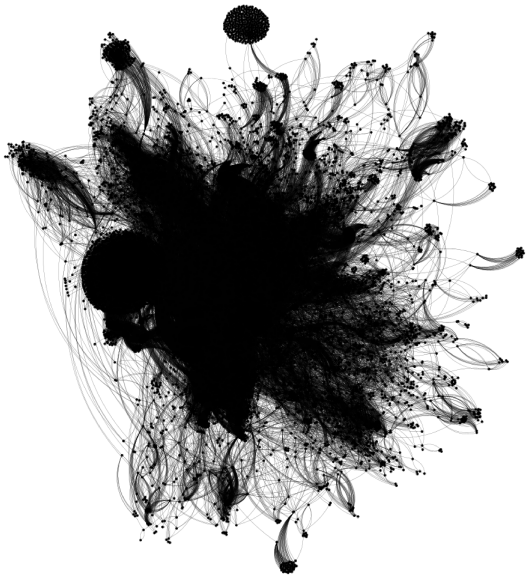


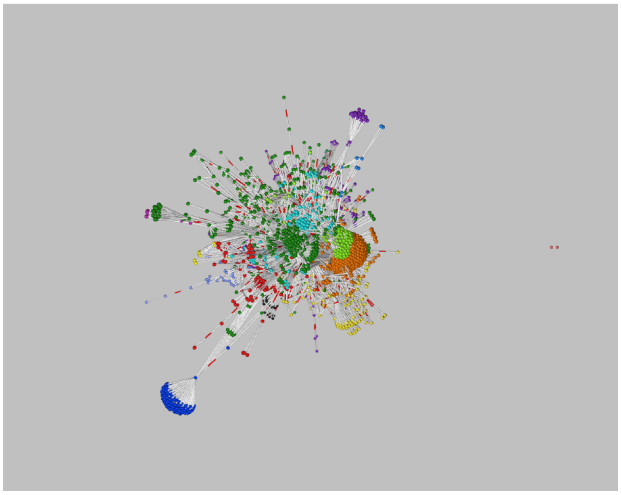
7.2 Squirrel





7.3 Crocodile





7.4 Directed Crocodile