Network Analysis
Epidemic Spreading

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

1.3 Roadmap

The main focus of this report is the diffusion of an epidemy within our graphs.

We will mainly analyze the differences in the results of our simulation when we change the parameters that affect the spreading behavior. In particular we will focus on these scenarios:

1. transmission probability increases,
2. recovery probability increases,
3. minimum number of rounds of infection changes,
4. initial infected number changes.

And we will also analyze what are the implications in choosing the initial spreaders:

1. with a random-based strategy,
2. with a centrality-based strategy.

Chapter 2

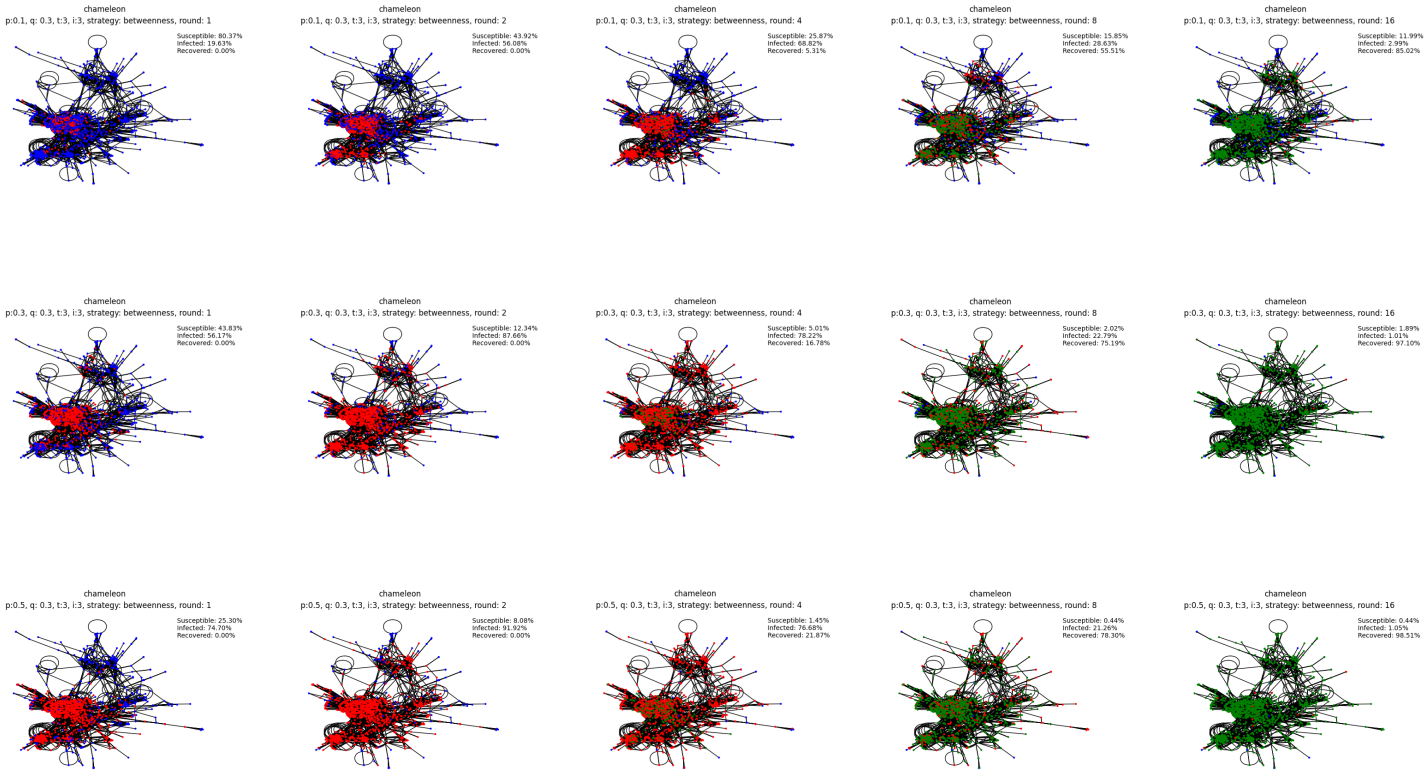
Parameters Impact

In this section, parameters impact on the spreading of the epidemy will be analyzed.

2.1 Transmission Probability

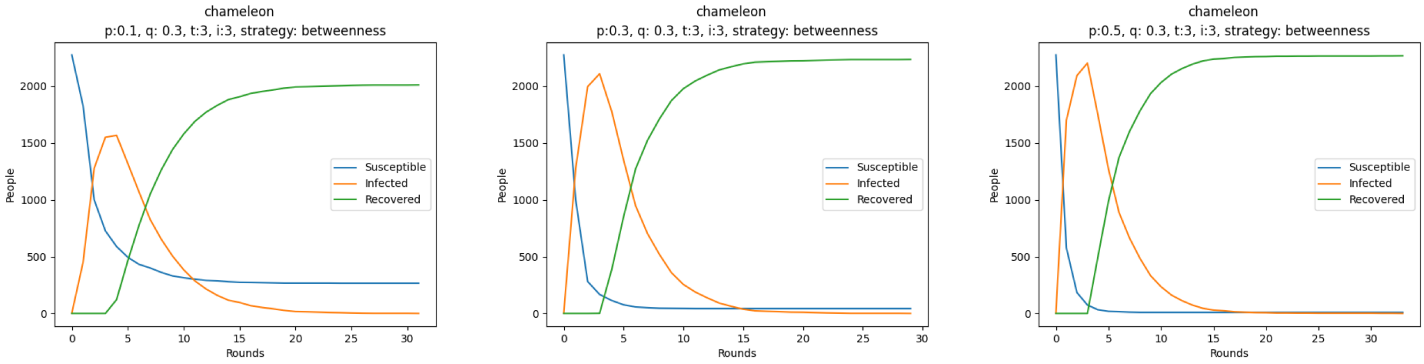
This parameter regulates the probability with which an individual infects each of its neighbors during each round spent as infected:

Let's consider one clear example, fixing the parameters to *betweenness strategy*, $\mu = 0.3$, $t = 3$, $i = 3$ and setting β to 0.1, 0.3, and 0.5 in the first, second and third row of the image below, respecitvely.



As the infection probability increases, the number of infected in rounds increases dramatically. So, **it has much more impact the change from 0.1 to 0.3 then the change from 0.3 to 0.5**.

We can consider now the SIR plots, indicating all the three sets changes over rounds.



We can also notice that there is a crucial change in the behaviour of the

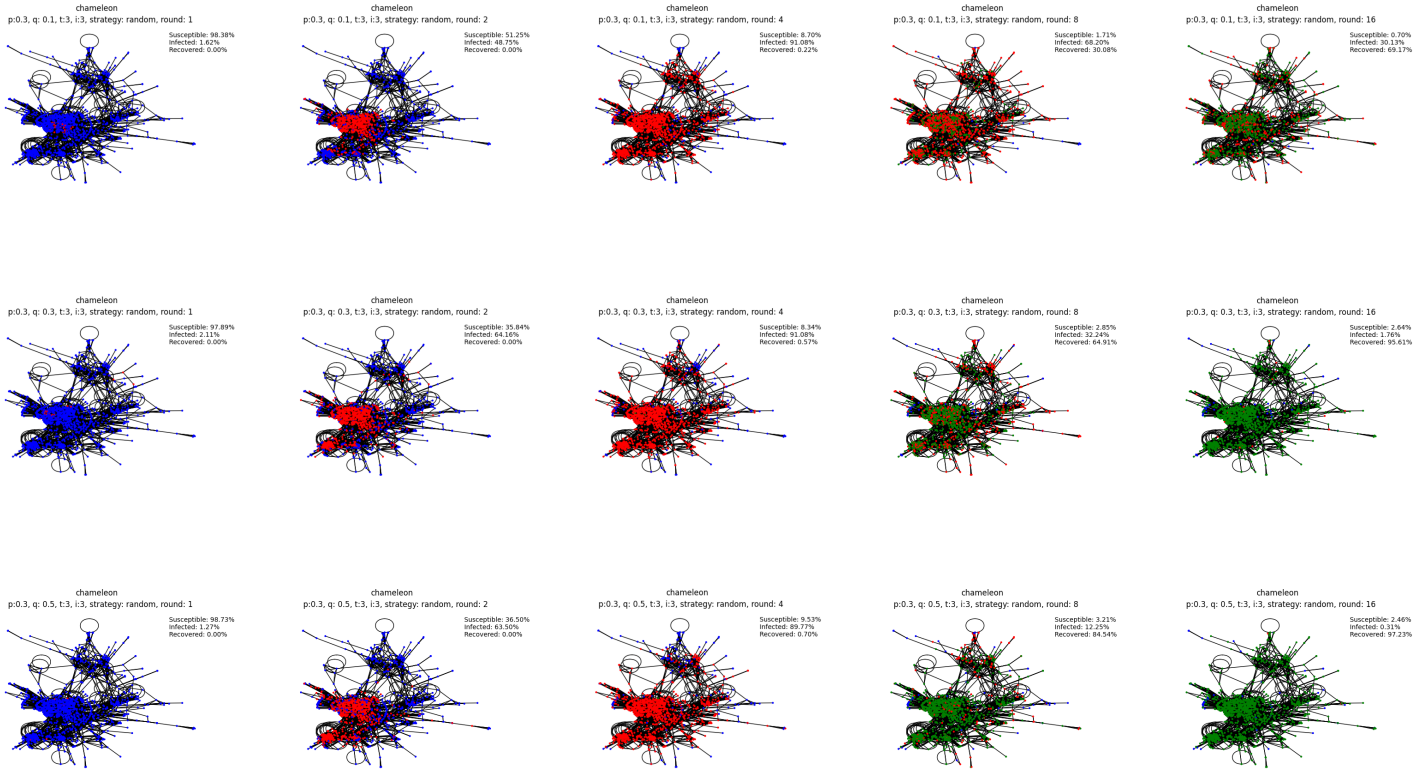
simulation **between 0.1 and 0.3**:

1. for 0.1, on the left, infected curve grows at a slower pace,
2. for 0.3 and 0.5, the curves representing the number of infected nodes have a higher peak. Furthermore, we can notice that there are a lot less susceptible nodes at the end of the simulation with these values for β

2.2 Recovery Probability

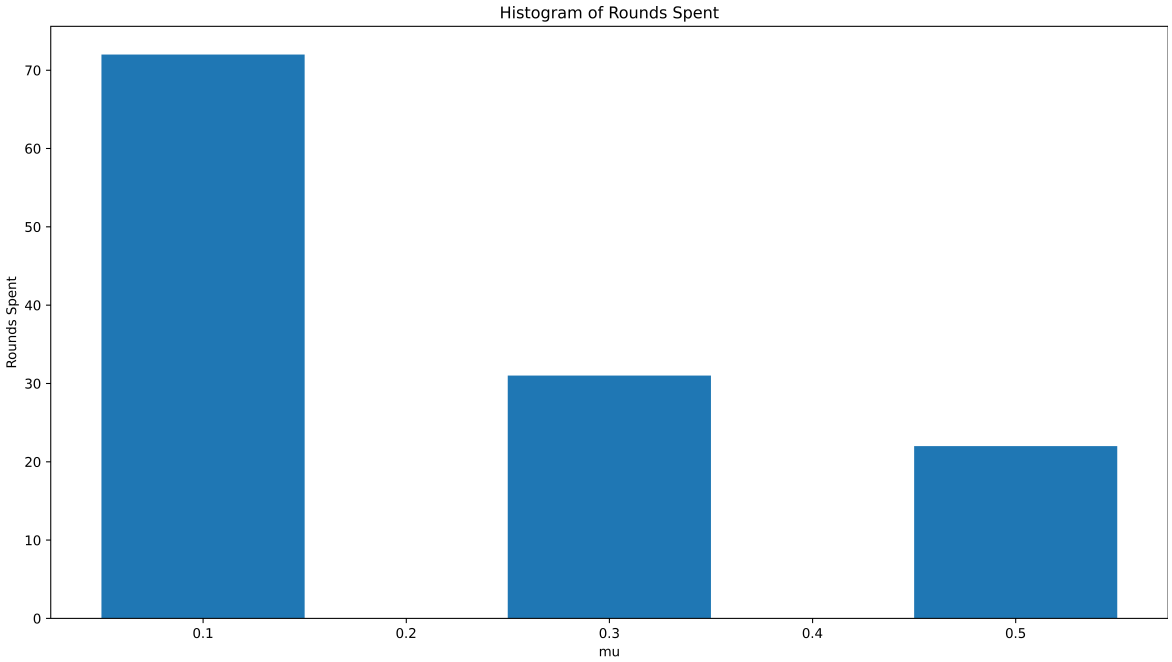
This parameter regulates the probability that each infected node has to recover during each round after the node has been infected for t rounds.

Let's consider one clear example, fixing the parameters to *random strategy*, $\beta = 0.3$, $t = 3$, $i = 3$, and setting μ to 0.1, 0.3, and 0.5 in the first, second, and third row of the image below, respecitvely.



μ acts as the inverse of β, and so also here we can notice that **the biggest difference can still be seen between $\mu = 0.1$ and $\mu = 0.3$**, albeit being a smaller difference, since the recovery time is still gated by the t parameter.

In this case we can also notice that the number of rounds needed to complete the simulation sharply decreases as μ decreases, as it can be seen from the plot below:

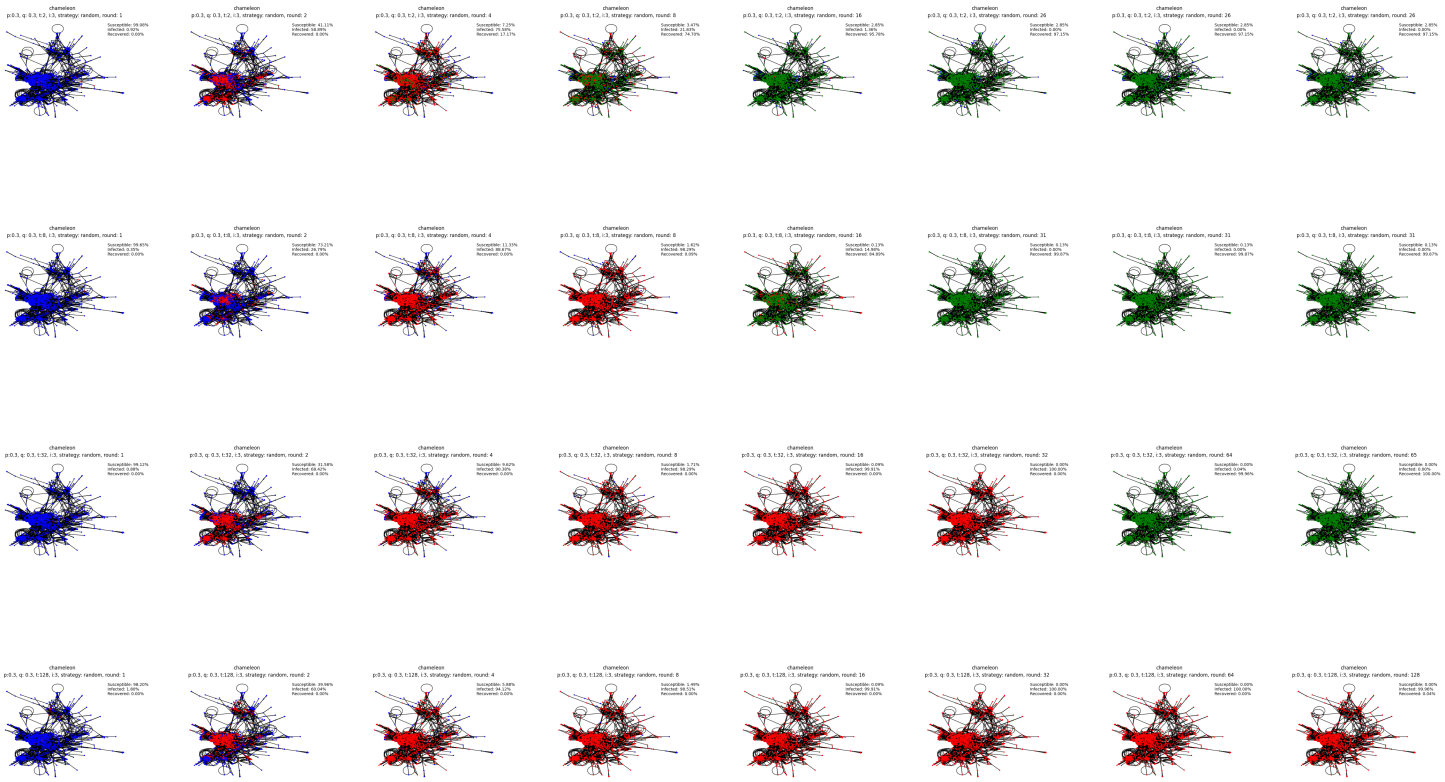


2.3 Minimum Rounds of Infection

This parameter regulates the minimum number of rounds that each node has to spend as infected.

In the plots below, we fixed the parameters to *random strategy*, $\beta = 0.3$, $\mu = 0.3$, $i = 3$ and setting t to $[2, 8, 32, 128]$.

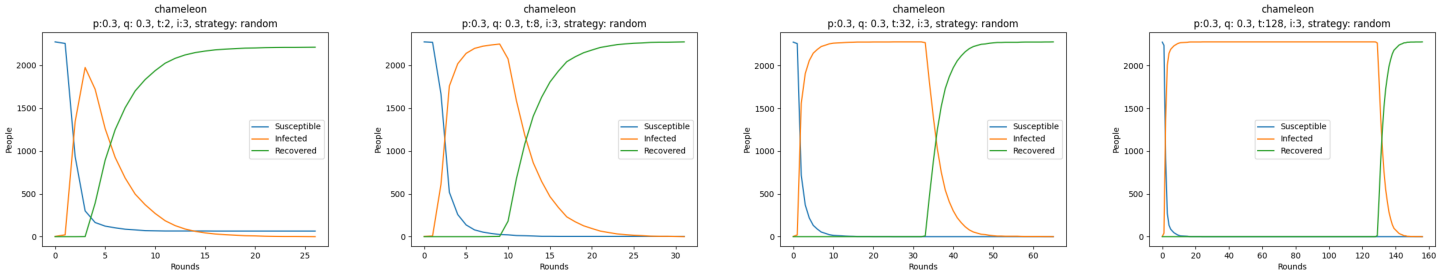
Note: when a simulation finishes before the round for the column, the last round of the simulation is shown.



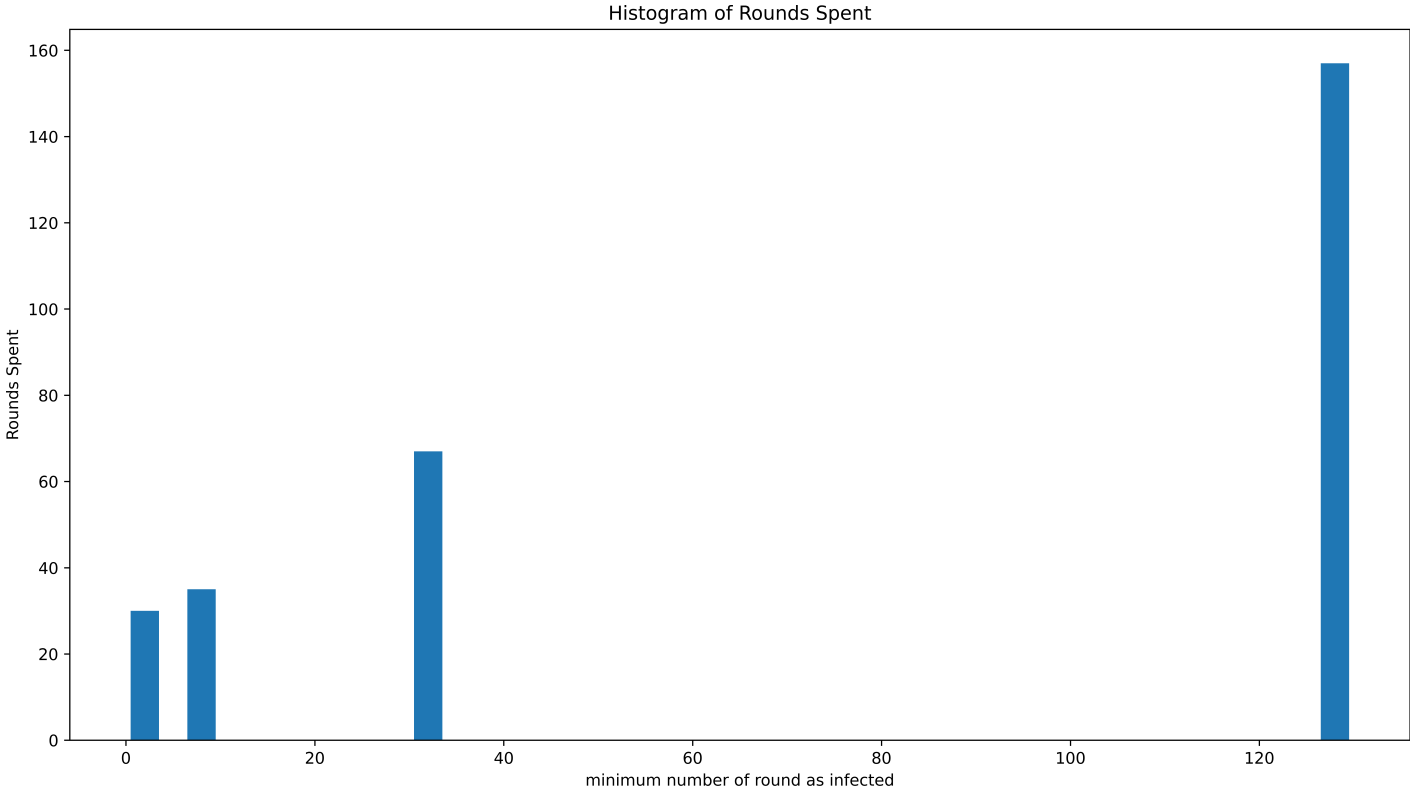
As we can easily notice from the plots above, the simulations take a higher amount of rounds to complete, since the max number of rounds is set to infinite, and thus the simulation waits for the epidemy to die out, which is gated by the minimum number spent as infected.

The same behaviour can be observed from the plots below, in which we can see how the plots further on the right show a wider peak of infected people.

Furthermore, observing the SIR plots we can notice that the epidemy is always able to reach all of the nodes in the network regardless of the suboptimal choice of the first infected nodes. This happens because the simulations are forced to run for a higher number of rounds, so nodes are exposed to infected neighbours for a larger amount of time.



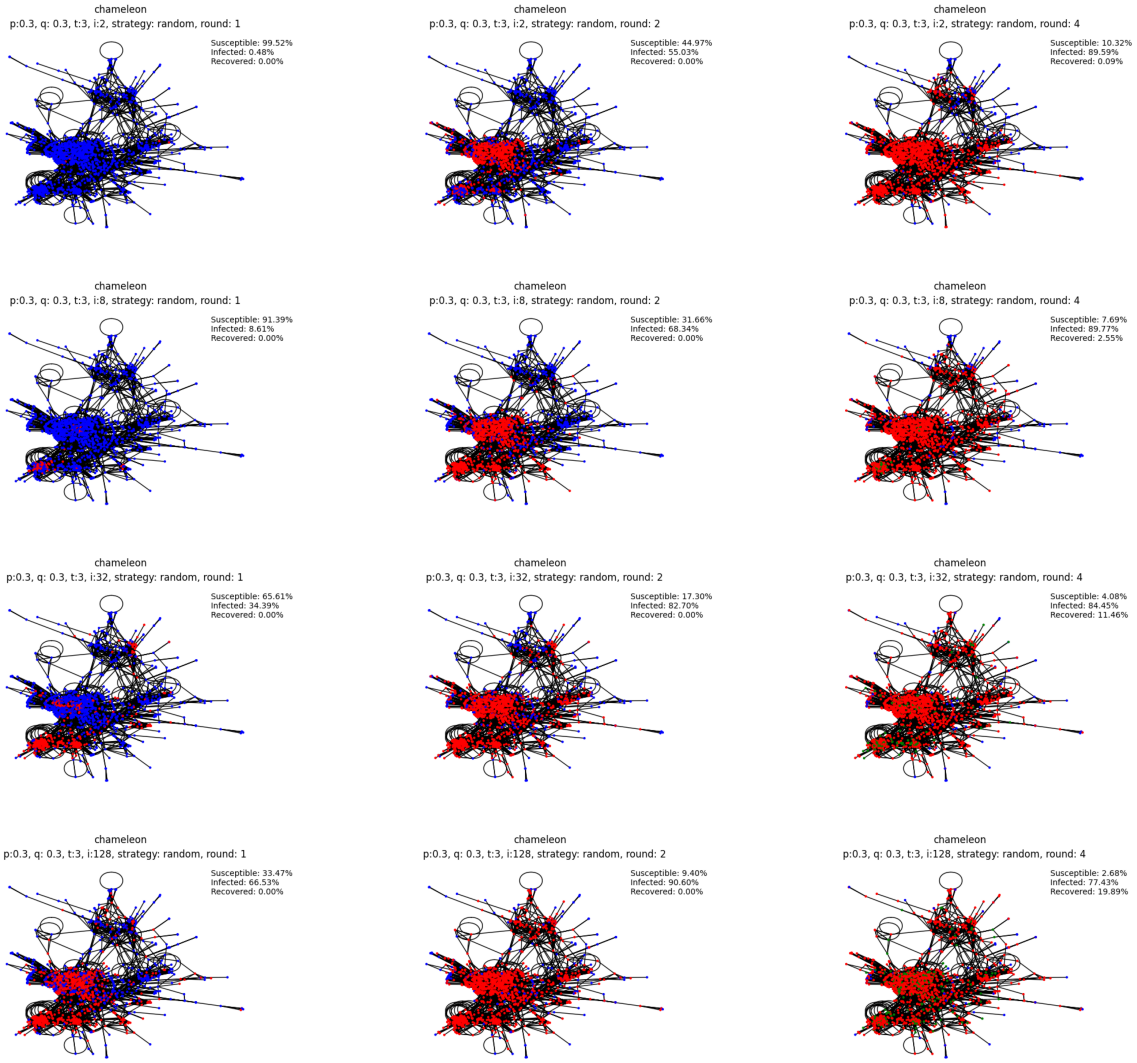
Lastly, in support of the last assertion, we can see that the number of rounds spent in the simulation increases almost linearly as t increases. As reported above, this happens because with higher values of t most of the rounds in the simulation are spent waiting for the minumum number of rounds to pass. During this period, the epidemy is still able to spread, while nodes are not able to recover, thus the nodes can "flip the infection coins" more and more, until the random factor is nullified.



2.4 Initial Infected Count

This parameter regulates the initial number of infected nodes. These nodes are **the only ones** chosen based on the attack strategy.

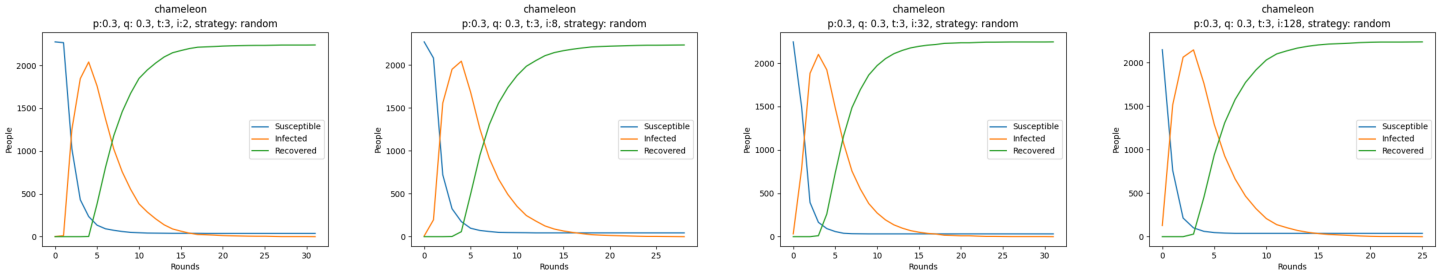
In the plots below, we fixed the parameters to *random strategy*, $\beta = 0.3$, $\mu = 0.3$, $t = 3$ and setting t to $[1, 2, 4]$.



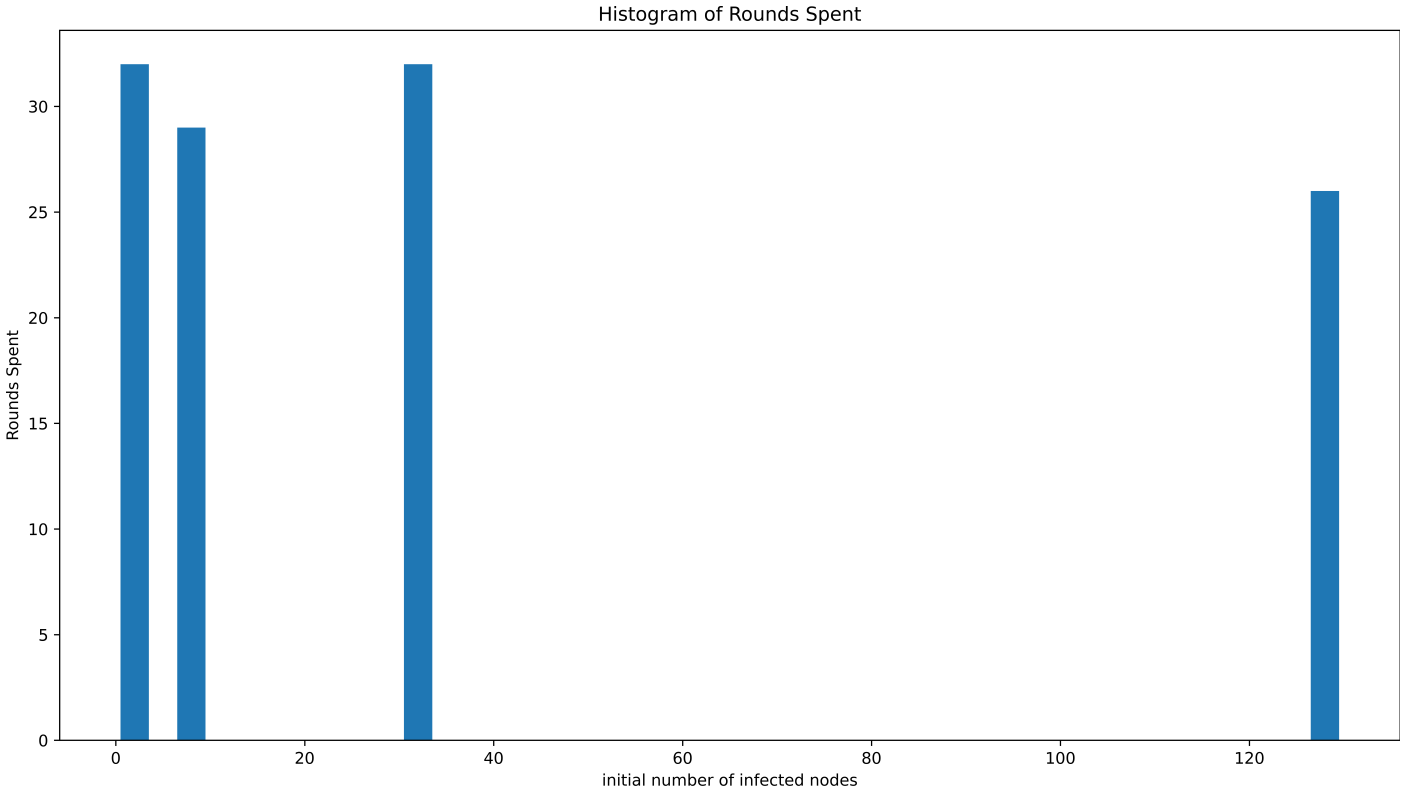
The images show how having a higher initial infected count basically results in starting with a headstart in rounds.

Then, after the initial rounds, the difference between simulations flattens

out. The only divergence we can notice in SIR plots is a slightly increase in the number of infected when i increases:



This parameter doesn't affect the final number of rounds that much since β is set to a somewhat high value. The difference would be much bigger with a lower value for β, since the simulation would need a higher number of rounds to reach the corresponding number of initial infected nodes.



Chapter 3

Strategies Comparison

In this analysis we will focus on the two main strategies used to choose the first infected nodes:

1. **Random Selection**,
2. **Target Selection**,

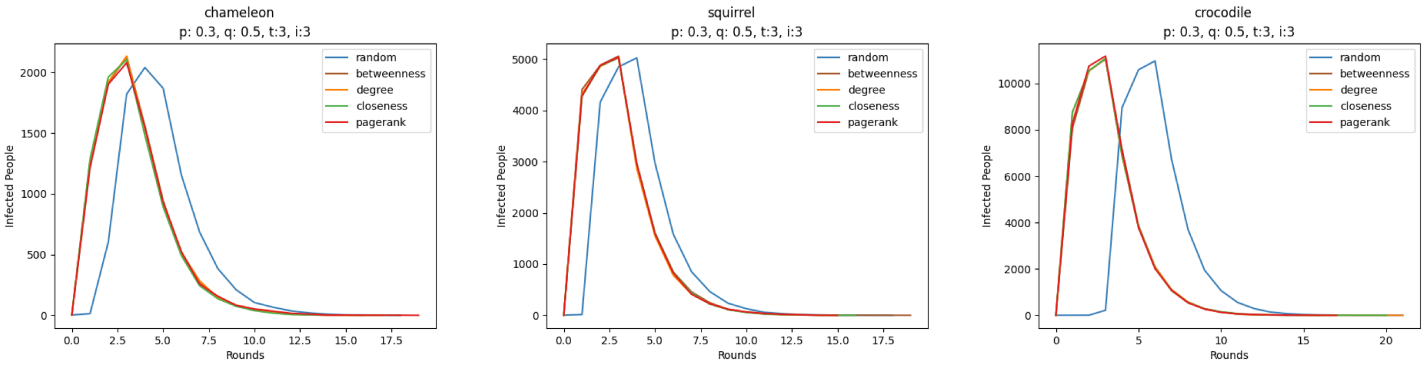
3.1 Expected Results

In the previous report, we were able to show how all the datasets represented scale-free network, with the scale-free regime becoming clearer as the number of nodes increased.

In the case of a targeted infection on (some of) the hubs in a scale-free network, we expect the epidemy to spread more rapidly, and this part of the report aims to prove this hypotesys.

3.2 Empirical Results

These are the results regarding the three datasets:



We can notice that all the centrality-based (Target) strategies works out in a similar way. On the other hand, we can clearly see that the random infection strategy has less impact, especially in the initial rounds of the simulation.

When comparing the strategies across the datasets, we can observe that there is a bigger "spreading speed" gap between random and the targeted strategies on the **Crocodile** dataset

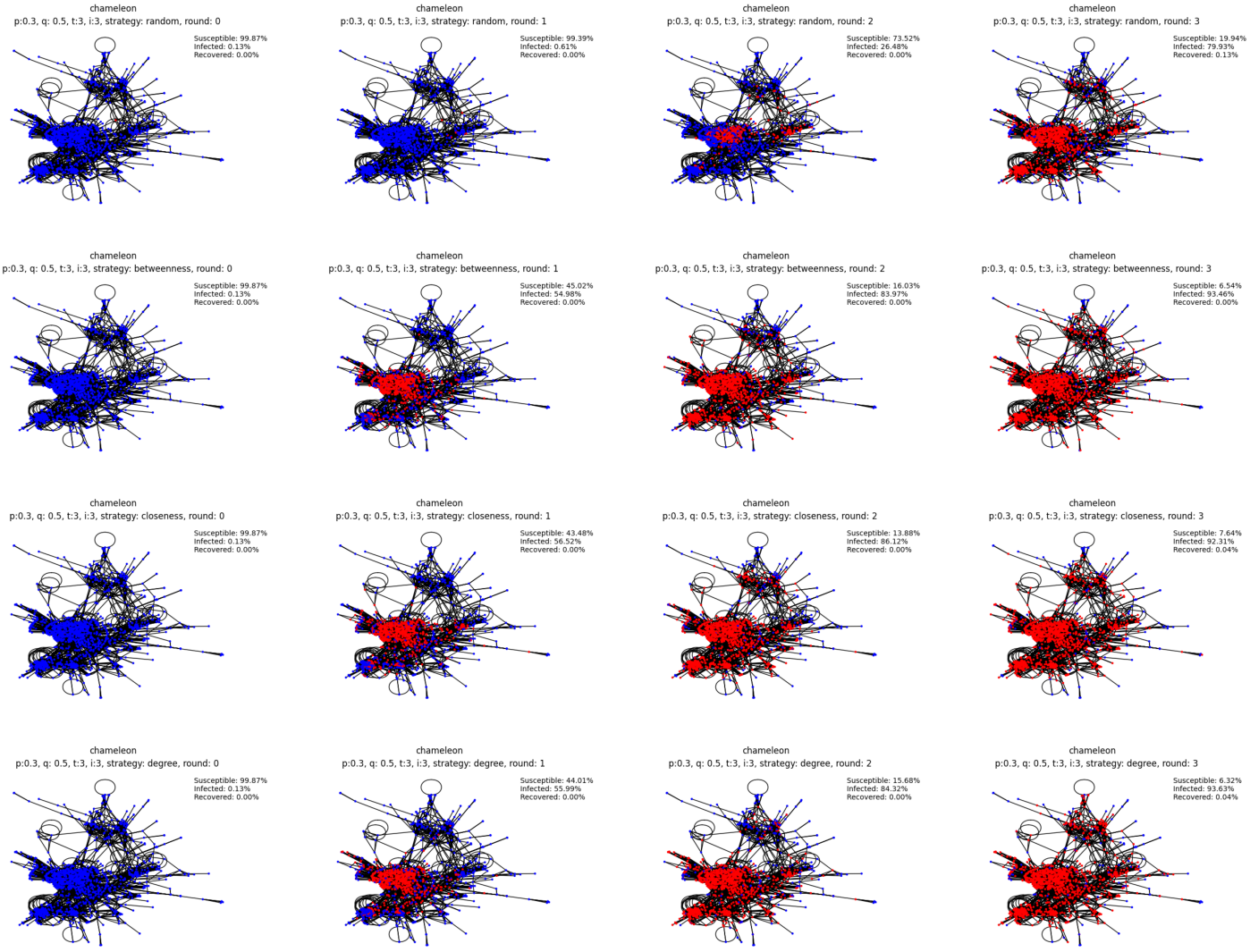
So, regarding strategies:

1. All of the datasets show approximately the same curve for all the centrality-based strategies,
2. **Crocodile** seems to be more resistant to the random-based strategy. This is to be expected, since it shows a clearer scale-free structure.

3.2.1 Two Examples Comparison

In this image, the dataset (Chameleon) and most of the parameters are fixed, with the only parameter changing being the strategy. From top to bottom, the listed strategies are:

1. random
2. betweenness
3. closeness
4. degree



It's clear to see that the infection propagates at a slower pace (it picks up speed later on into the simulation, when more central nodes have probably been reached) with respect to other strategies.

Chapter 4

Impact of Communities

When we use a SIR models to simulate the spreading of information or diseases on a small-world or ultra-small world network, we should observe that:

1. the epidemic initially spreads among communities,
2. weak ties start spreading the epidemy from one community to another later on into the simulation.

The aim of this chapter is to prove this hypotesys.

4.1 Core Example

All of the following observations, images, and pieces of information can be found or deducted by looking at **this explainer video**. The video shows the difference in the spreading of the epidemy among communities. On the left, we can see a random-based infection, while on the right we can see a betweenness-based infection strategy. All of the other parameters are the same.

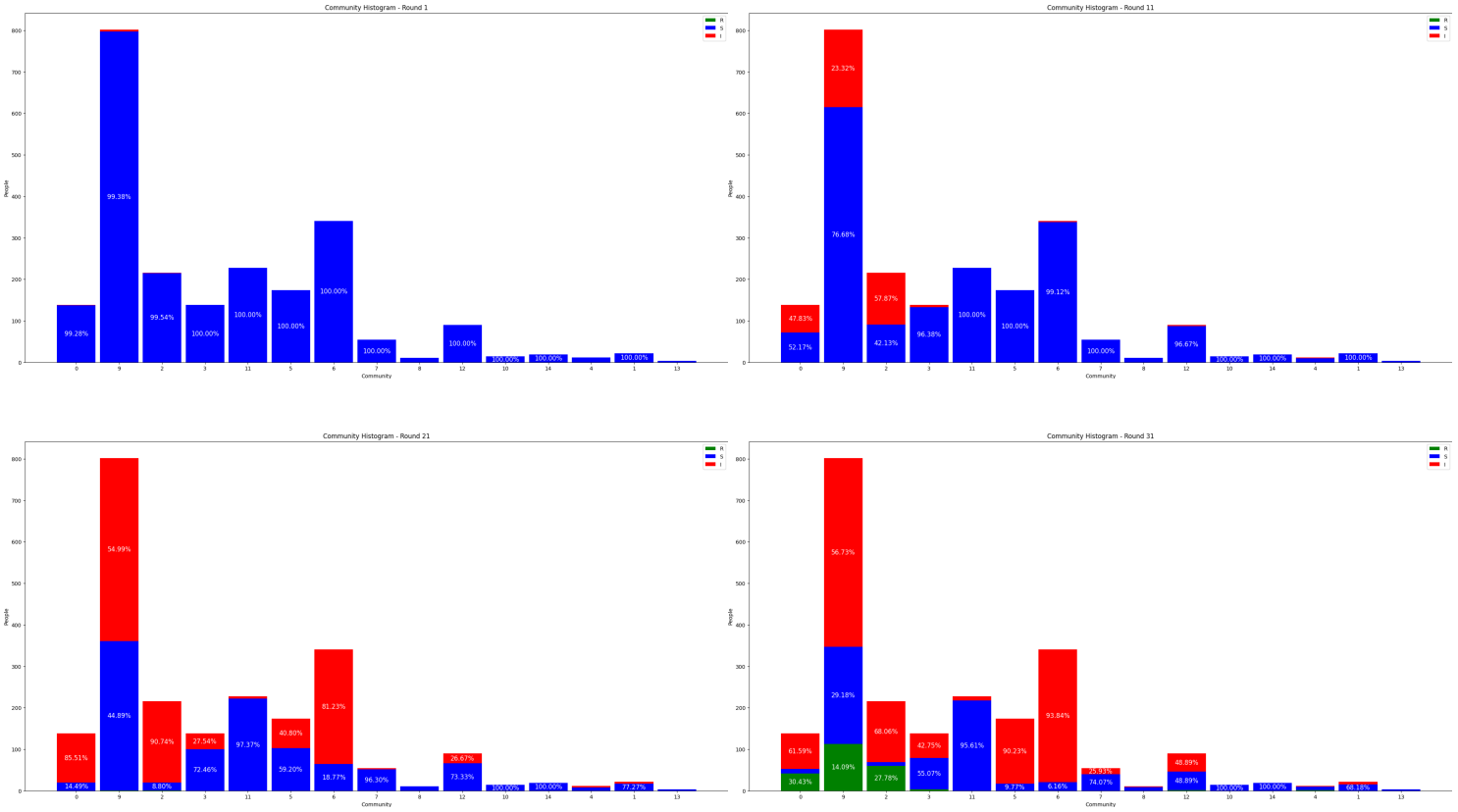
The video clearly shows that:

1. the spreading starts from communities and then spreads in other communities using weak ties,
2. the evidence that a betweenness-based infection is spreads faster than a random-based infection.

4.2 Spreading among communities

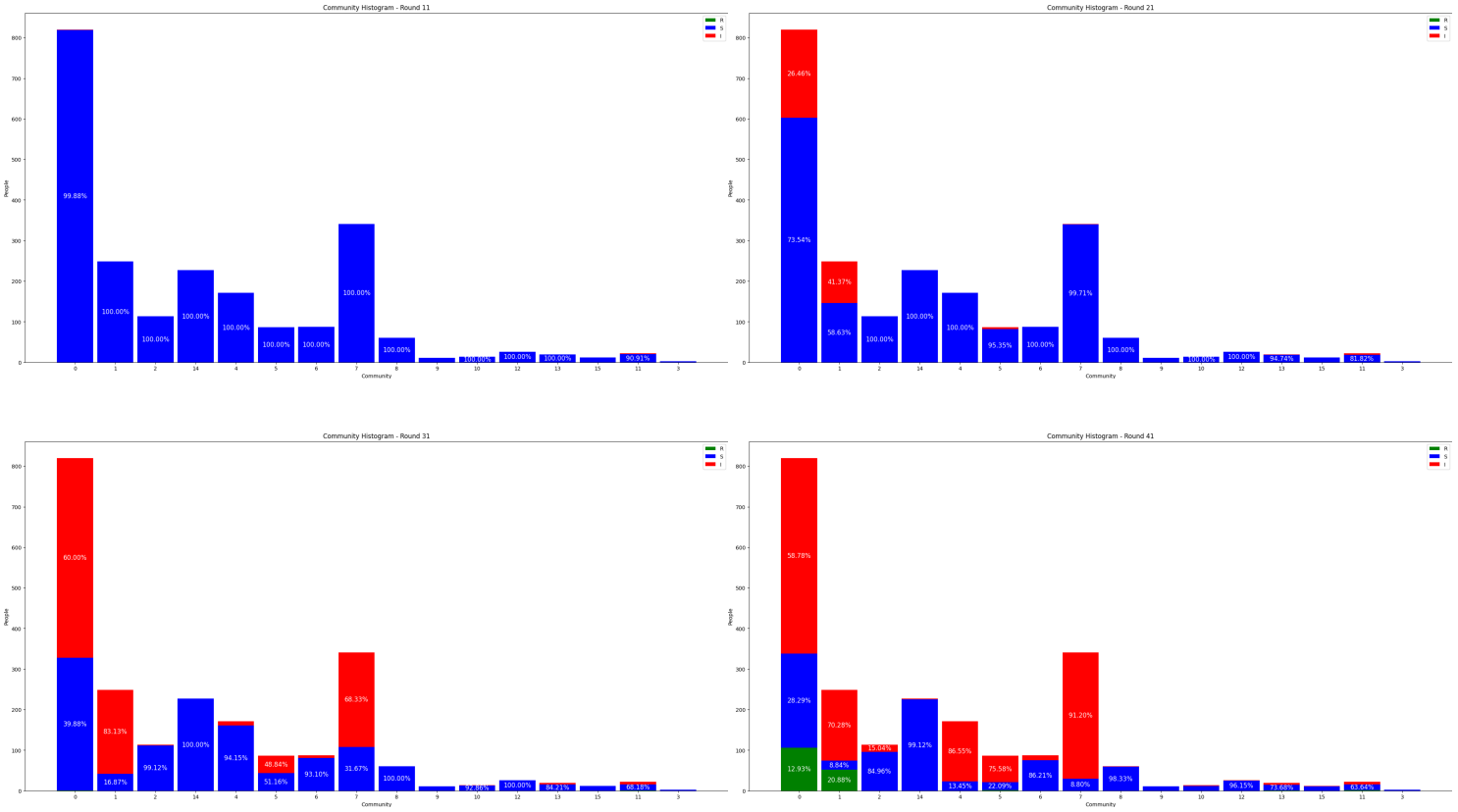
This section is a further analysis on the key points taken from the comparison video.

In this first plot, we can see rounds $1, 11, 21, 31$ of a betweenness-based infection:



We can clearly notice that the infection in round 11 is spread **only among 3 communities**. Then it starts spreading among the others during the course of the simulation.

As comparison, in this second plot, we can see rounds $11, 21, 31, 41$ of a random-based infection. Its important to notice that we are showing the situation 10 rounds later in the simulation (in respect to the first plot) for all images:

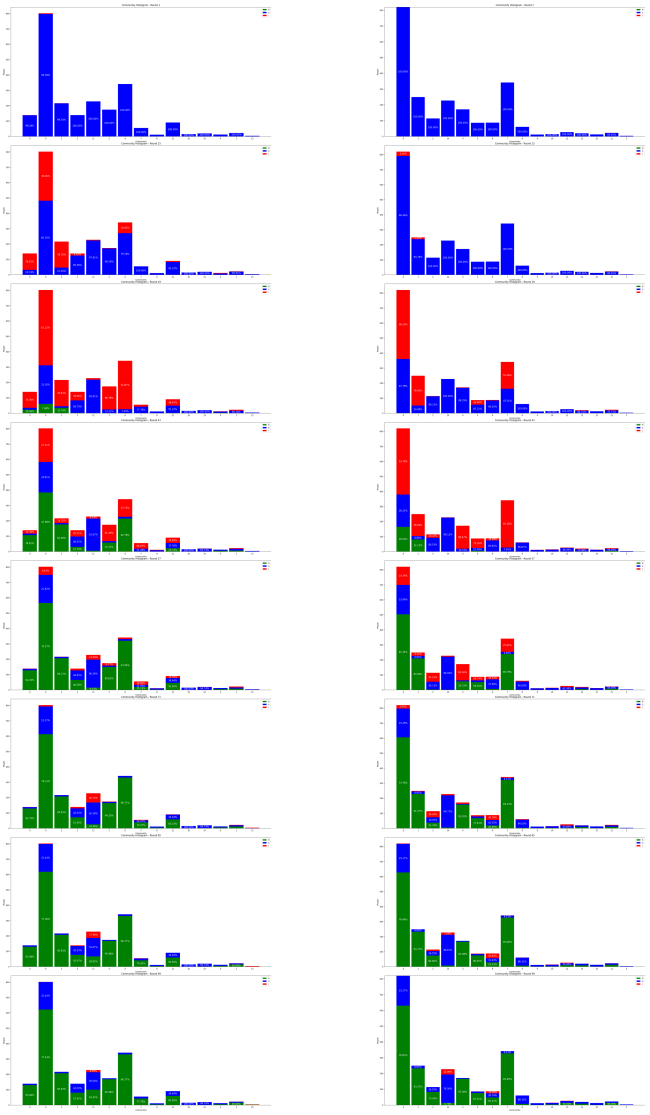


The **infection starts its spreading at a lower pace**. In these plots we can also notice that the spreading starts among some communities and then propagates to the others.

4.3 Random vs. Centrality First Infected Nodes

In this final figure we have:

1. a betweenness-based infection of initial nodes on the left.
2. a random-based infection of initial nodes on the right.



This is further evidence of the faster spread if we pick the initial nodes from the pool of the most central ones with respect to one of the various centralities or page-rank score.