INFECTIOUS DISEASE SPREAD

João Ribeiro (77209) | Ricardo Rei (78047) | Raquel Casteleiro (82027)

Complex Networks 2017/2018

INTRODUCTION

A complex network is a mathematical structure that models how a population of entities behave with one another. This structure consists on a set of nodes, which represent the entities, and edges, which represent the relationships between nodes. Networks can be applied to almost anything in most study areas.

In this work, network science is used to model a human contact network related to the propagation of diseases in a population. It is based on a study done on Stanford University [1] (California, USA), where all credit for the previous analysis is due.

Infectious diseases are usually passed via droplets during close proximity interactions and thus the pandemic spread of an infectious disease poses a big threat to society in several ways.

This work aims to understand how a disease might spread across a population, so that us humans can understand how to fight it, by studying on a low-scale controlled environment such as a high school. Each individual was given a proximity sensor that registered its close proximity interactions. These devices have a radius of 3 meters and had a coverage of 94% of the total school's population interactions.

"Schools are particularly vulnerable to infectious disease spread because of the high frequency of close proximity interactions" [2]

During our work, we developed some code in Python[3], using the NerworkX[4], package that can be found in our github: https://github.com/RicardoRei/Complex-Networks-17-18

THE DATASET - NETWORK AND NODES

The dataset used consists of a single-day recording obtained from an American high school. The network built is a weighted undirected network.

Nodes represent individuals (i.e., people) and edges represent close proximity interactions between said individuals.

There are 788 total nodes (655 students, 73 teachers, 55 staff and 5 other people). The edge representation is detailed in the next session.

The dataset can be found on the original study's web page.[5]

EDGES – CONTACT REPRESENTATION MODELS

Each edge represents close proximity interactions between two individuals. An interaction between two individuals is defined by a continuous sequence of close proximity records (CPRs), stored on the individual's device. Therefore, there is the subtle issue of how to represent relevant details such as the duration and number of the interactions between two individuals.

The dataset already solved this issue for us by recording interactions in four different strategies:

1) Add-then-chop
2) Chop-Then-Add
3) Chop-then-count
4) Just-Chop

The first 3 make use of the "minimum duration" parameter that defines that the minimum duration (in CPRs) for an interaction must be set and the last one makes use of the "drop-off" parameter that defines the minimum CPR gap to be filled (allows you to assume that the dataset might be missing CPRs).

The first strategy first adds all CPRs registered by the interactions between two individuals to create the weight of the edge between the two and then applies the minimum duration parameter, i.e. doesn't consider edges with a weight

less that the value of the minimum duration parameter. The second strategy first applies the minimum duration parameter to all interactions between two individuals and then adds the remaining interactions in order to create the weight between the edges of the 2. The third strategy returns the number of interactions ignoring the number of CPRs in each interaction. Finally, the last strategy uses the drop-off parameter in order to fill “gaps” between interactions, e.g. if drop-off = 1 and there is an interaction between A and B that registered 3 CPRs in time steps 1, 2 and 3 and there is another interaction between A and B that registered only one CPRs in time step 5 the gap between these 2 interactions will be filled in order to create a bigger interaction with 5 CPRs in time steps 1, 2, 3, 4 and 5.

The dataset’s edges lists all assume that minimum duration parameter value is 1 and the drop-off is 0.

For our analysis we used the Add-Then-Chop model because in our opinion it’s the simplest one to understand and represents well the difference between small contacts and big contacts.

Network Metrics

Degree

Figure 1 - Weighted Degree distribution

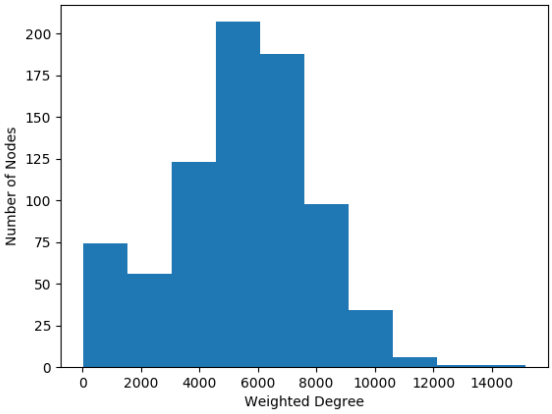


Figure 2 - Cumulative Weighted Degree distribution

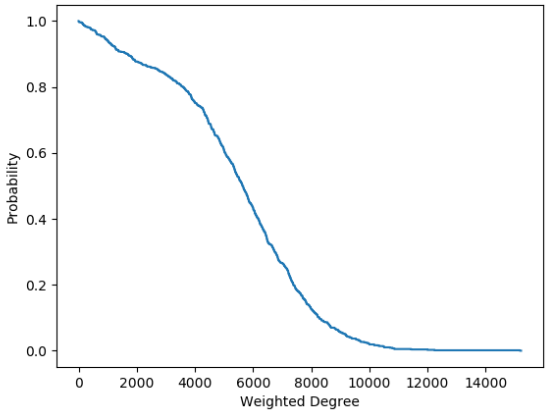


Figure 3 - Cumulative Distribution of Contacts time

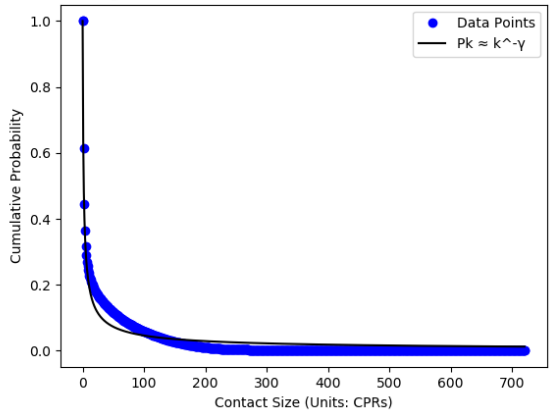
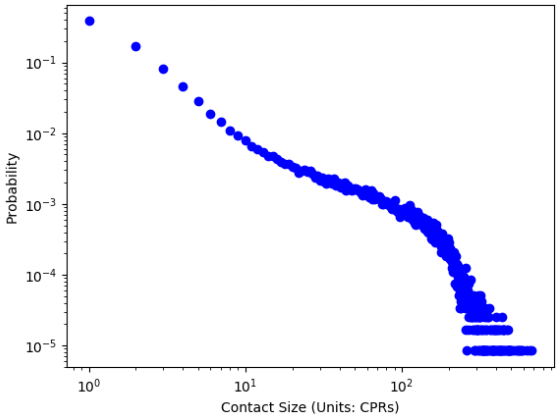


Figure 4 - Degree Probability



The degree of a node tells how connected a node is.

In the disease spreading context, the degree of a node will represent how many contacts a person had with others during a typical school day.

The average degree of the network, if we ignore the weights, is approximately 300 which means that in a normal school day each person has in average 300 contacts. Note that not every contacts have the same duration and, in terms of an infectious disease spreading, if individual A is infected, the longer the contact between A and another individual B, the higher is the risk of B being infected.

With that said it's important to look for this metric considering weights in order to understand the size of the contacts. The weighted degree is based, not only on the number of edges, but is also pondered by the weights of each node. Figure 1 shows the histogram of the weighted degree and figure 2 shows the cumulative weighted degree distribution. We can see by these figures that the great majority of individuals have between 3000-9000 CPRs with an average of 5425 CPRs per day, recall that a CPRs is approximately 20s which is a lot of time in a 3m range from other individuals. Note that this value of a CPR considers overlapped time steps, e.g. if individual A stayed for 1 hour with 30

people the degree of the node representing A will be something like 5425 ((5425 * 20s)/60s/60m = 30h) but in reality person A only stayed 1 hour in contact with other individuals.

Another interesting thing is to see the distribution of the contact's size or in other words the distributions of the weights in the graph. Figure 3 and 4 show that the number of small contacts is a lot bigger than the number of longer contacts and inclusively the relation between the probability of having a contact with size k and the size of the contact itself is a power law with gamma = 0.667.

Average Path Length

The average path length considering no weights measures the average distance between 2 pairs of individuals.

In context of disease spreading these metric tells how far someone is in average from having a contact with an infected person.

The average path length is 1.6219 considering no weights. This means that in average the distance between 2 pairs of individuals is another person and "half". Since there are no "half" people the intuition is that person A had contact with almost half of the school directly (path length = 1), and to the remaining population, A had contact with someone that had contact with them (path length = 2) so in average we get 1.5 = (0.5 * 1 + 0.5*2).

Considering weights, the intuition is the same but we use the weight to "bring closer" individuals where the contact length was higher and vice-versa. The average path length with weights is 1.87 which is not much different from the average path length without weights because the great majority of contacts only have size 1 (as shown in figure 3).

Diameter

The diameter represents the distance between the most remote individuals in the network.

The diameter of this network is 3. In case of a disease spreading, it means that if individual A is one of the extremes, A will have to infect 2 other people in order to have a chance to infect the other extreme B and by the time B is infected the probability of reversing the disease spreading is low.

Looking to the average path length and the diameter of this network, we can easily see that these network presents typical **small-world properties**.

Clustering Coefficient

"Mind your friends' friends."

The clustering coefficient tells the extent to which the nodes group themselves in a network. This means that a network with a higher clustering coefficient is more clustered than a network with a lower clustering coefficient.

When it comes to the context of disease spreading, this information may tell us which parts of the network may be infected faster than others. If a node in a very clustered group gets infected, then spreading a disease across the group itself may be faster than if the node infected was in a not very clustered group.

In this study, the **most clustered** individual was a **student** (id 26), with a clustering coefficient of approximately 0.047 and the **least clustered** individual was a **staff member** (id 375), with a clustering coefficient of approximately 0.0014.

The average closeness centrality was approximately 0.0056.

We can see that all values are pretty low, and we can derive that this population, when it comes to close proximity interactions, is not very divided into clusters (probably because being in someone's proximity isn't the same as having social interactions with him/her (and where clusters would be naturally found). This seems positive, seeing that having a low clustering coefficient is something that may help when it comes to the containment of a disease. However, if we were to fight the spread by focusing on individuals, then students would be the best bet (seeing as they have the largest clustering coefficients). If one were to dig deeper out of curiosity in how a disease spreads through clustered populations, there is a paper online about the topic.[3]

Node Metrics

Degree Centrality

"How popular are you?"

The degree centrality of a node is based on the number of nodes it is connected to and therefore represents the risk or probability of that node catching what is spreading through the network.

In case of disease spreading, it shows the risk of someone getting hit by the disease (i.e., the higher the degree centrality the higher the risk) and, if trying to contain its spreading, the people with the most degree centrality should be the first ones to be protected.

In this population, the individual with the **higher degree centrality** (approximately 0.6696) was a **student** (id 171) and the **lower degree centrality** (approximately 0.0051) was a **staff** member (id 376). In the results that can be found by running our code, it can be seen that the **average of the students' degree centrality** (approximately 0.4203) is the **highest** one, being almost the double of the teachers' (approximately 0.2238) which comes in second place.

With these results, it's easy to understand that the class most at risk of being infected by a disease spreading throughout the network is the students' and, so, they should be the ones to receive protection first.

Eigenvector Centrality

"My friends are better than yours"

The eigenvector centrality measures the influence of a node in the network, based on that node's connections (i.e. the more influent the nodes it's connected to, the more influent is the node).

In case of disease spreading, the node with the highest eigenvector centrality, which means the most influent node, would be the best "patient zero" if the objective is to spread the disease the fastest way possible. On the other, hand, if trying to prevent the spreading of the disease, an epidemy that starts with this node would be much more difficult to contain.

In this population, and not taking into account the weights of the connections, the individual with the higher eigenvector centrality (approximately 0.0585) is a student (id 171). When taking into account the weights of the edges, the node with the higher eigenvector centrality changes (approximately 0.1078) to another student (id 520). By running our code, some other results will be found and it can be seen that the average of the different groups' eigenvector centrality doesn't change much wether we take or not into account the weights of the connections, but the **student's** have, in both cases, the **higher average eigenvector centrality** (approximately 0.0371 and 0.0351, for a not weighted and a weighted network respectively).

It's interesting to note that the most popular individual (see Degree Centrality above), is also the most influent one in a non-weighted network and it's a student (id 171), but when the network becomes weighted it isn't the most influent one anymore. We can determine that the students are the most influent group of the school and that they should be the priority when trying to prevent the spreading of a disease.

Closeness Centrality

"How close are you?"

Closeness centrality of a node tells how centrally positioned in the network the node is (i.e., the higher the value, the more central in the network the node is).

When it comes to the context of disease spreading, this information may tell which individuals would be the best "patients zero". The closer an individual is to the entire network (i.e., the greater its closeness centrality) the more likely it is to aggravate the spreading of a disease were he's the patient zero.

In this study, the **most central** individual was a **student** (id 171), with closeness centrality value approximately **0.750**. The **least central** individual was a **staff member** (id 376), with a closeness centrality value approximately **0.371**. These results make sense seeing as students are more involved in social gatherings than any other school's group of individuals. School's staff's work on the other hand doesn't involve much social interaction.

The average closeness centrality was approximately 0.621.

These results show us that when containing a disease, the best way is to go for the individuals with largest closeness centrality, in this case, mostly students.

Betweenness Centrality

"How do you influence the flow of information?"

The betweenness centrality of a node tells us how much a node connects other groups of nodes (i.e., if it was removed, then the groups it connected wouldn't be directly connected anymore).

When it comes to disease spreading, it tells us which individuals might help spread the disease from group to group (and therefore their containment is crucial).

In this study, the **most in-between** individual was a **staff member** (id 16), with betweenness centrality of **0.006958**. The **least in-between** individual was also a **staff member** (id 267), with betweenness centrality of zero (0.0). There was an average betweenness centrality of 0.001103 (very low).

These results show that when it comes to close proximity interactions, individuals in a school interact as if they were a large group (there are no split groups connected by single nodes). Therefore, in the context of disease spreading this is something that wouldn't help against the containment of the disease itself.

Conclusion

In summary, after analysing these data we can conclude that the network revealed small-world properties with homogenous contact structure in which short interactions dictate. We could also conclude that, when trying to contain or start an epidemy, the students should be the focus, since they proved, troughout our analysis, to be the most popular, most influent and in general the most connected group.

The characteristics of the network helped us realize how fast an infectious disease can spread over a population so common as a school and how easily things can scale to other populations. It's important to keep in mind that this dataset has some limitations specially because it doesn't measure other contamination routes like surfaces, etc. Furthermore, different pathogens might have different minimum contamination requirements than a 3-meter radius range.

The next step in our work will be to present actual disease spreading simulations, possible immunization strategies and maybe differentiating between different diseases.

References

[1] http://sing.stanford.edu/flu

[2] http://sing.stanford.edu/pubs/PNAS-2010-1009094108.pdf

[3] https://www.python.org/downloads/release/python-362/

[4] https://networkx.github.io/documentation/stable/index.html

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2817154/