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 behaves 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 the 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 it’s close proximity interactions. These devices have a radius of 3 meters and had a coverage of 94% of the total school’s interactions.

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

**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 persons). The edge representation is detailed in the next session.

The dataset can be found on the original’s study’s web page[1]

**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 a 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 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 resisted 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.

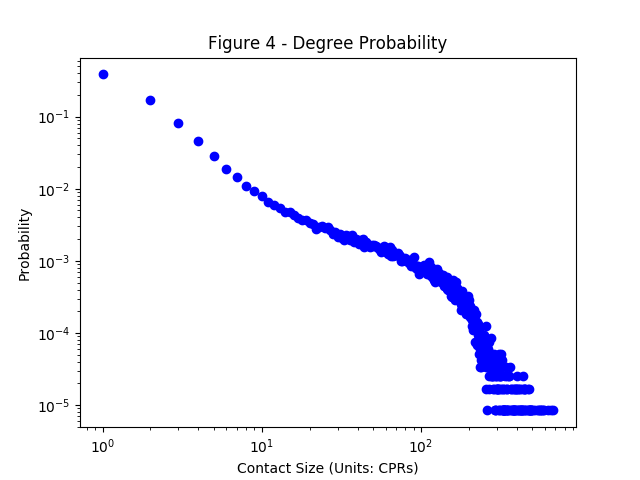
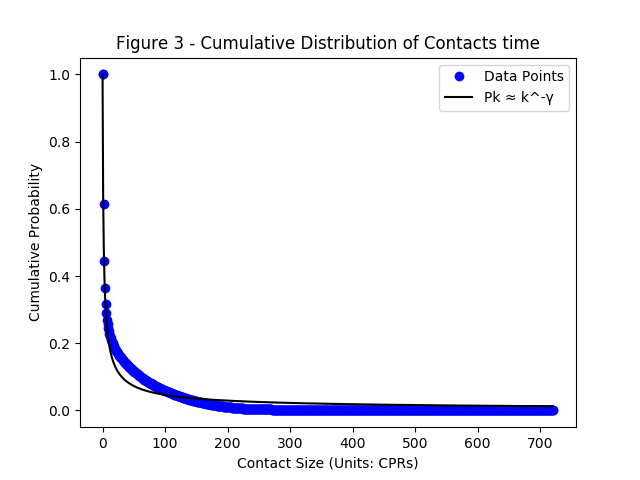
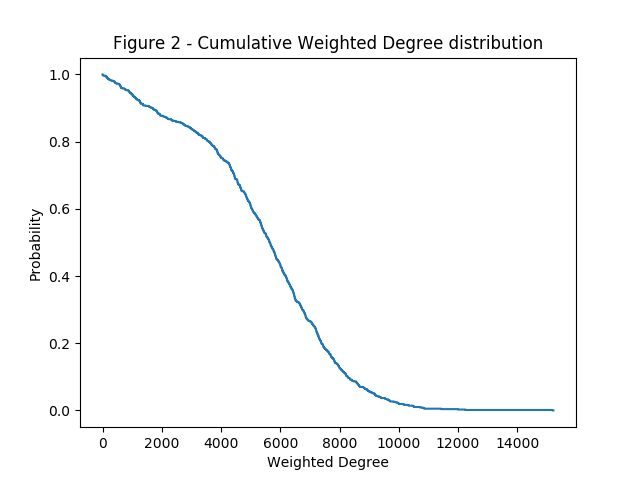
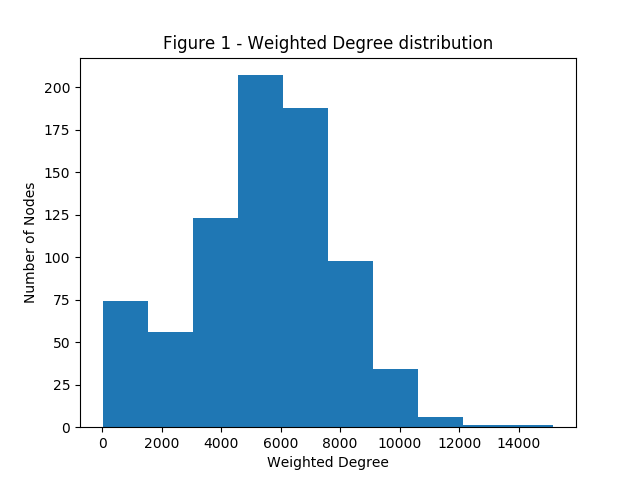
Colocamos algum exemplo?

The dataset 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 is the simplest one to understand and represents well the difference between small contacts and big contacts

**Network Metrics**

**Degree**



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 another individual B the higher is the risk of B is infected.

Being that said it's important to look for the degree metric considering weights in order to understand the size of the contacts. The weighted degree it’s based not only on the number of edges but 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 4000-8000 CPRs with an average of 5425 CPRs per day, recall that an 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 persons 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 contacts size or in other words the distributions of the weights in the graph. Figure 3 and 4 shows that the number of small contacts it’s 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 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 is no “half” persons the intuition is that person A had contact with half of the school directly (path length = 1), and to the other half, A had contact with someone that had contact (path length = 2) so in average we get 1.5 = (0.5 \* 1 + 0.5\*2).

Considering weights, the average path length 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 of this network is 3. These metric represents the distance between the most remote individuals in the network and in the context of a disease spreading it means that if individual A is one of the extremes, A will have to infect 3 other persons 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**

*“The extent to which the nodes group themselves in a network.”*

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 spread, 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 group wasn’t very clustered.

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, it’s 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 itself. However, if we were to fight the spread by focusing on individuals, then students would be the better bet (seeing as they have the largest clustering coefficients). We found a paper online that discussed the spread of infectious disease through clustered populations[3]

**Node Metrics**

[RC] **Degree Centrality**

ToDo

[RC] **Eigenvector Centrality**

ToDo

**Closeness Centrality**

*“How close an individual is to all other nodes of the network?”*

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 spread, this information may tells which individuals would be the best “patients zero”. The closer an individual is to the entire network (i.e., greater closeness centrality it has), the more likely it is to aggravate the spreading of a disease were he 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 class of school individual. School staff 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**

*“The extent to which a node lies on paths between other groups of nodes.”*

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**

ToDo

**References**

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

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

[3] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2817154/>