CS-E5740 Complex Networks, Answers to exercise set 7

Sara Cabodi, Student number: 784287

November 11, 2019

Compile with `pdflatex ex_template.tex`

Problem 1

a) In this exercise I performed a weighted network analysis using a social network data set describing private messaging in a Facebook-like web-page.
The plot below shows the complementary cumulative distribution (1-CDF) for node degree k, node strength s and link weight w.

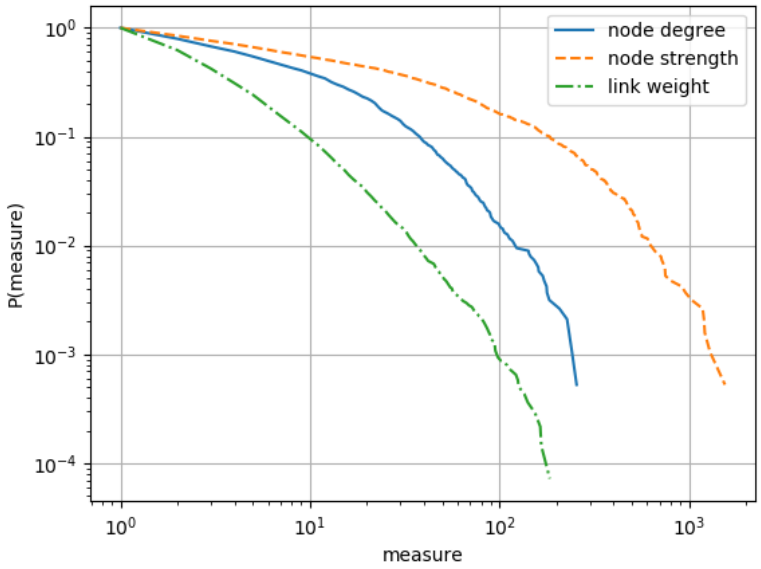
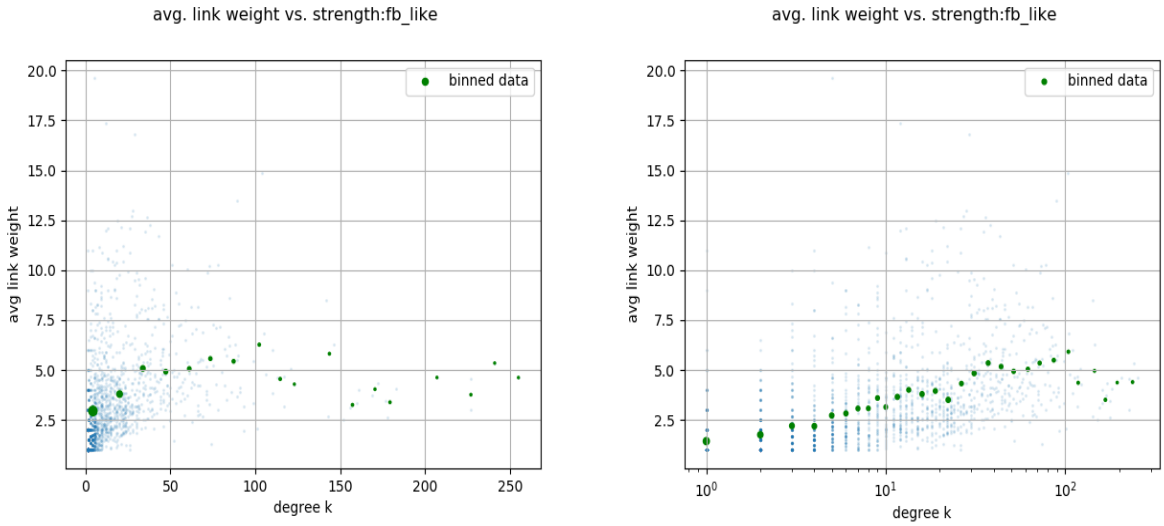


Figure 1: 1-CDF on Facebook-like web-page network

The distributions are not Gaussian. They are not even power laws, because they would have been straight line if they were. Although, their initial behaviour is similar to power laws up to 10 on the x-axis, they all tend to decrease faster than the power laws with higher x values.
Based on the plots, I can roughly estimate the 90th percentiles of the three distributions as follows:

- degree $\rightarrow 4{*}10^1$
- strength $\rightarrow 2{*}10^2$
- weight $\rightarrow 10^1$

b-c) The following scatter plots show $\langle w \rangle$ as a function of k. These plots aim at studying how the average link weight per node $\langle w \rangle = \frac{s}{k}$ behaves as a function of the node degree. To make the relationship between $\langle w \rangle$ and k more visible, the plots include a bin-averaged version of the scatter plots with a number of bins equal to 20.



(a) Linear axes

(b) Logarithmic axes

d) I think that the logarithmic approach suits better in order to represent $\langle w \rangle$ as a function of k. This is because it better shows a trend, especially for lower degrees. It has an increasing trend up to degrees around $k = 20$, which is in contrast with the usual behaviour of social networks. For larger degrees (more than 20), there is not a definite trend (proven by several trials with different values of the size of bins), but the binned data tend to spread both over and under the average link weight of 5. One possible explanation could be that for higher values of degree, fewer data are available and so the average link weight is less uniform. Moreover, since these are real data, the lack of samples increases the impact of noise in them. Another possibility could be that for low degrees, links have a more uniform average weight

than those with higher k, that tend to spread unevenly.
In social network, the trend is usually the opposite ($\langle w \rangle$ decreases as a function of the degree) so this specific case of the online community for students at University of California is in contrast with that behaviour. One possible explanation could be that nodes (students) with few interactions (low degree) have also low weight on them because maybe they prefer other ways to communicate or they tent to stay a bit isolated, while students that have a lot of interactions tent to have significant (high average weight) ones too.

e) The following scatter plot shows the overlaps as a function of link weight. As in c), there is also a bin-averaged version of the plot. For this purpose I chose the logarithmic approach with a number of bins equal to 20.

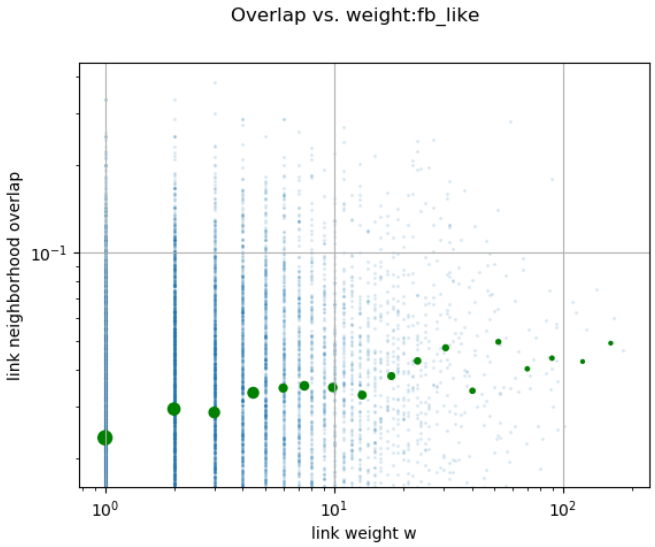


Figure 3: Logarithmic plot

The plot shows that the link neighborhood overlap can be seen as an increasing function of link weight, which agrees with the Granovetter hypothesis. Although it can be quite accordant with the hypothesis, with more trials with a growing number of bins, the binned data points tend to spread after a $\langle w \rangle$ around 10^1. This behaviour confirms the hypothesis for low average link weights (up to 10) and tend to discard it for higher values of $\langle w \rangle$. This could maybe due to the fact that there are less links with higher weight, so the link neighborhood overlap tend to be more randomized and it does not follow a uniform trend.

Problem 2

a) Considering the network describing the US Air Traffic between 14th and 23rd December 2008, here are some useful basic network properties:

- Number of network nodes $N = 279$
- Number of network links $L = 2088$
- Density of the network $D = 0.05384$
- Network diameter $d = 4$
- Average clustering coefficient $C = 0.6465$

b) The full network is represented below with all links on top of the USA map.

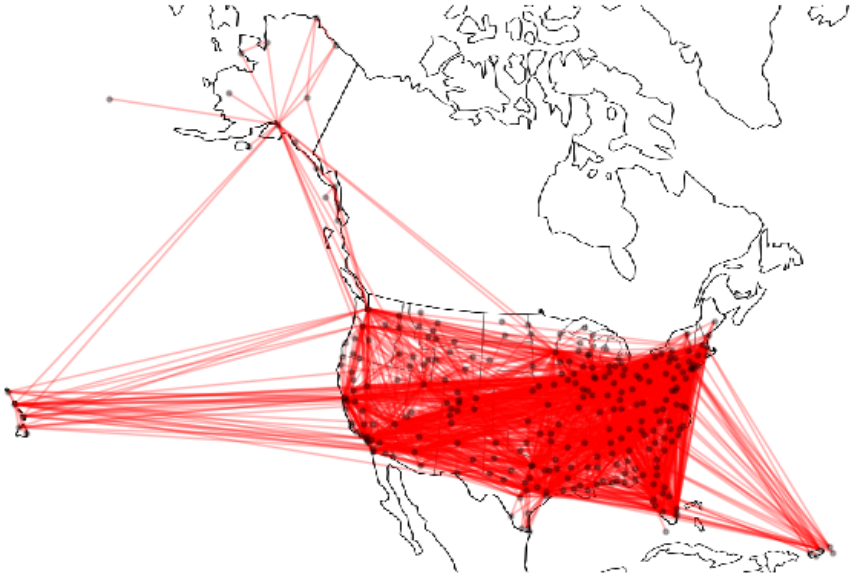
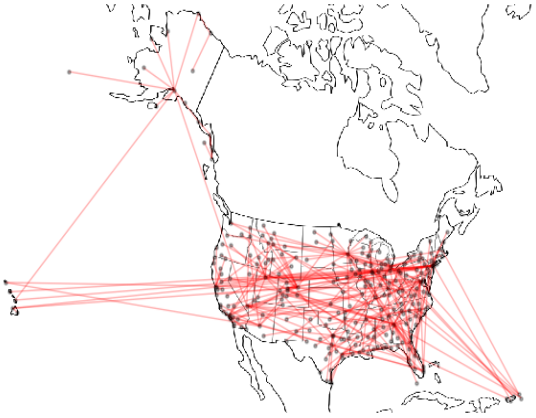
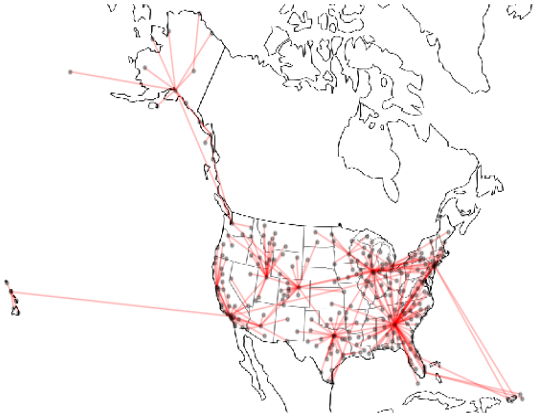


Figure 4: USA air traffic network

c) The computation of both the maximal and minimal spanning trees (MST) of the network are represented below.



(a) Minimal spanning tree



(b) Maximal spanning tree

The minimal spanning tree aims at minimizing the overall sum of edge weights, while the maximal maximizes it. In the case considered, where weights represent the number of flights, it is reasonable to consider that a link with heavy weight is the most convenient/popular terms of air traffic organization. So, the maximal spanning tree for Hawaii's connections is given by a single link connecting Honolulu to the mainland (Los Angeles) and other local flights/links to connect other airports in the Hawaii islands. The minimum spanning tree considers less popular/used connections, that, for instance, directly connect minor Hawaiian airports to Alaska and other US cities.

If I would like to understand the overall organization of the air traffic in the US, I would first use the maximal spanning tree because I feel it better highlights more the important airports/hubs. As already noticed, heavy weight links are those ones with more flights, so the best solution in order to organize the air traffic in the US would be to start by analyzing those ones. In other terms, the maximum spanning tree would be the best initial graph, limited to 278 links, in order to maximize the air traffic.

d) The network obtained by taking only the strongest $M = 278$ links into account, where M is the number of links in the MST, is visualized in the next picture.

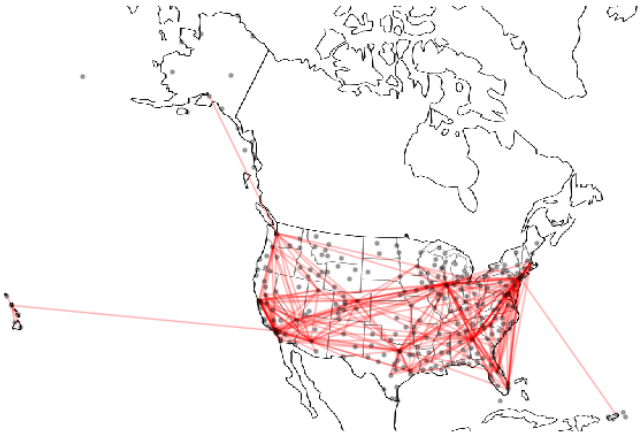


Figure 6: USA air traffic network threshold

The thresholded network shares 97 links with the maximal spanning tree. It is known from graph theory that a minimal/maximal spanning tree is not given by a set of (individually) minimal/maximal weighted edges: this is due to an optimality criterion that is global instead of a local one. Given this number (97 out of 278) and the visualizations, the simple thresholding does not yield a similar network as the maximum spanning tree. Furthermore, the two graphs highlight different aspects of the original network: the maximal spanning tree stresses the importance of certain USA hubs, while the thresholded network highlights the important connections. As a matter of fact, the graph actually looks like an intermediate one between the minimal and the maximal spanning tree; resemblance with the maximal spanning tree is higher for more isolated subgraphs, such as Hawaii islands, whereas the graph is closer to the minimal spanning tree in the mainland, where route redundancy is dominant with respect to simple reachability. A last interesting consideration is that the thresholded network is not connected (whereas the spanning trees are), as can be seen by many isolated nodes (peripheral cities with fewer connections) in the northern and mountain states, as well as in Alaska.