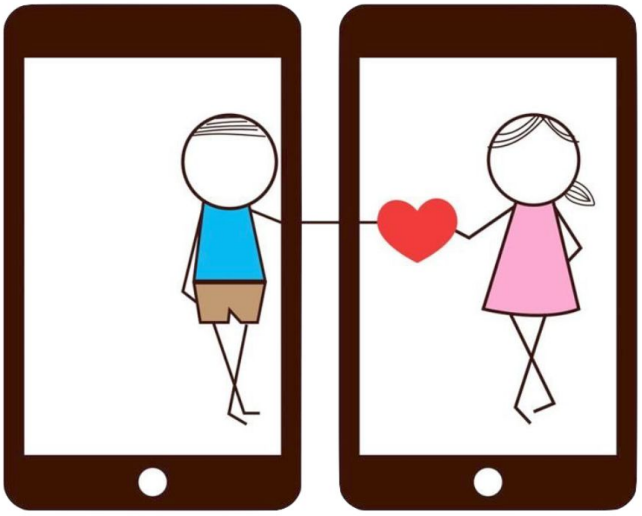
Rating Profiles

Teemu Koivisto 014211393
Tomáš Vopat 015148847
Github



Project Description

Our initial aim was to analyze the users' ratings of other users on a dating website. We were interested in finding out if the ratings or node degrees were somehow interlinked and were there some potential patterns in how the users behaved.

The dataset we used was from a Czech dating website called LibimSeTi from 2006 which consisted of two CSV files 250 MBs and 2 MBs each. They contained 17 million edges and 230 thousand nodes as a directed graph where the rating, integer between 1-10, from one user to another was the weight of the edge. For each node, there was a categorical value gender which was either Female, Male or Unknown.

The analysis

A preliminary goal was to see if the different genders behaved differently - were the ratings, node degrees or other features different for males and females. Then move on to further subset the data into different groups based on e.g. their sexuality or their out/in-degree. For correlation analysis between different features we used linear and logistic regression.

For graph properties, we used PageRank and triangle counting to find their distributions and the most prominent nodes. We also extracted a portion of the graph and transformed it into an undirected graph to analyze the basic community structure and calculate clustering coefficient.

Because our dataset was quite large, the analysis proved to be difficult with the full dataset. The enormous size prevented us from performing

many of the graph algorithms. We also had to redo many processing steps when the computation proved to be too much for the computer.
We wrote all the code to a Jupyter notebook using the standard Python libraries such as Numpy, Pandas, Scikit-learn and NetworkX. For more performant processing, we found out that GraphLab handled many of the more computation-intensive algorithms a lot better than NetworkX. For some of the visualizations, we used Gephi, which worked with moderately sized graphs (10-100 thousand edges).

The common adage of preprocessing and general hassle being 80% of the data scientist's work proved to be right yet again here.

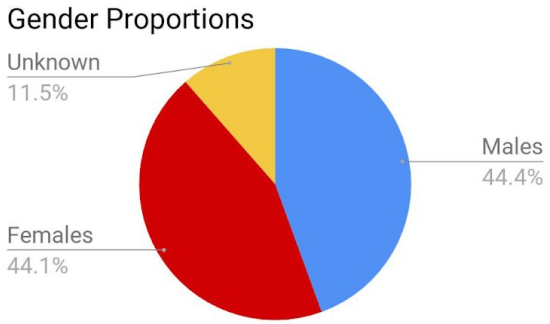
Preprocessing

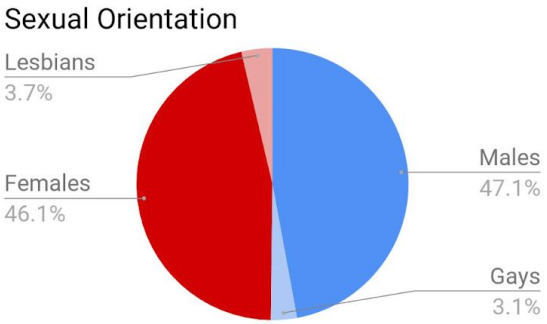
As mentioned, we had 17 million edges and roughly 200 thousand nodes. Luckily all of it was in a proper form without missing values so we were able to load it with Pandas without further issues. Yet the manipulation of the dataframes was not as trivial as one would imagine, and it took some effort to massage the data into a form that we deemed fit.

We condensed the 17 million edges into a another dataframe with 80 thousand rows, averaging the ratings mean and received ratings mean along with the incoming and outgoing degree counts. With this dataset, we did most of our statistical analysis.

Gender distributions

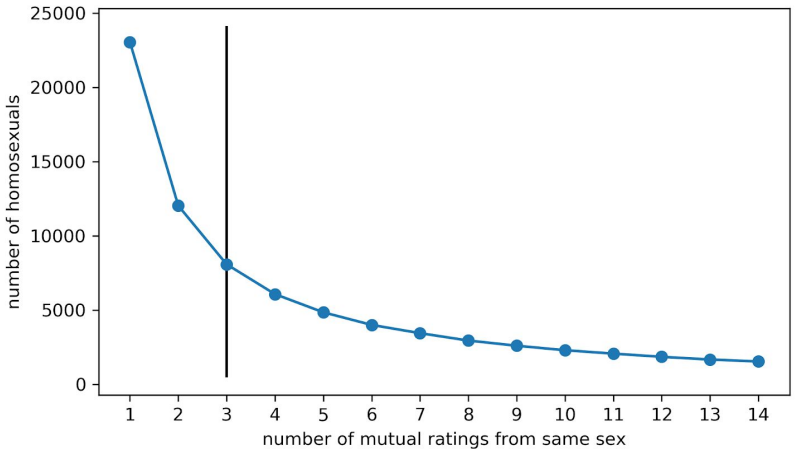
The first analysis we did was the gender distribution plots with their average values. Somewhat surprisingly, the genders are equally presented (gender proportion in dating sites is usually skewed towards male-heavy). Another peculiarity was the "Unknown" gender with their 11% of the gender distribution. They were simply users that did not want their gender to be public. For some reason, this also meant that they received far fewer ratings than normal users, possibly because of the lack of other information besides the gender.





We were also interested in finding out if we could determine the sexuality of the users based on their rating preferences. Meaning that if a user rates and receives ratings from their gender, can we assume them to be non-heterosexual.

To recognize these users, we created two undirected graphs (one for males and one for females), where we kept only the reciprocal edges. But since labelling all of these users as gays/lesbians/bisexuals would be too imprecise, we set the degree threshold to 3 (according to the following figure - there is an elbow).



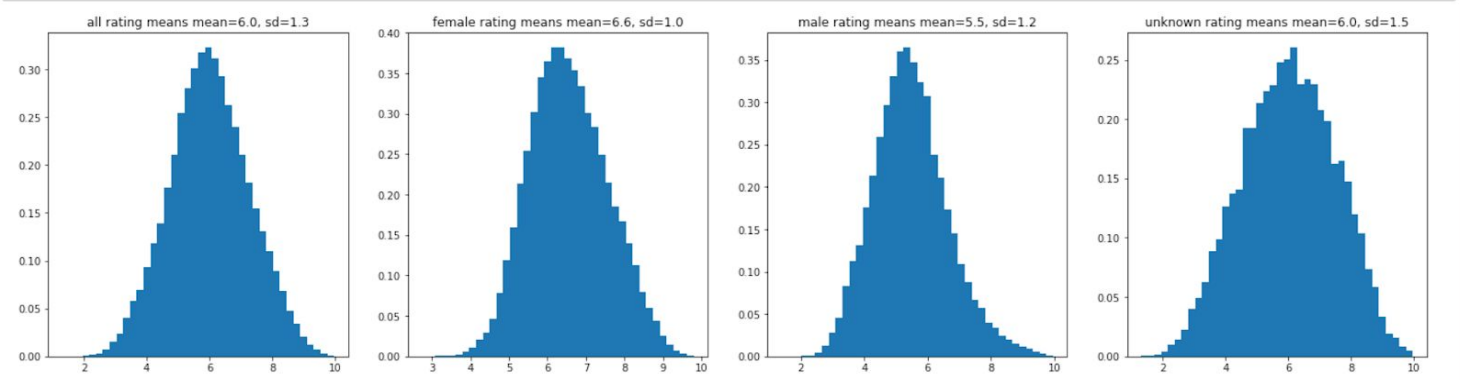
Some of those users were without any incoming ratings which we promptly removed. The resulting users were with more than 20 outgoing ratings and with male ratio (proportion of males in outgoing ratings) to 0.7 and 0.75 for males and females respectively. This threshold is a bit strict, but we wanted to avoid false-positives.

This percentage of lesbians, gays and bisexuals (LGB) seemed to be somewhat on line with the real distribution of said genders in population

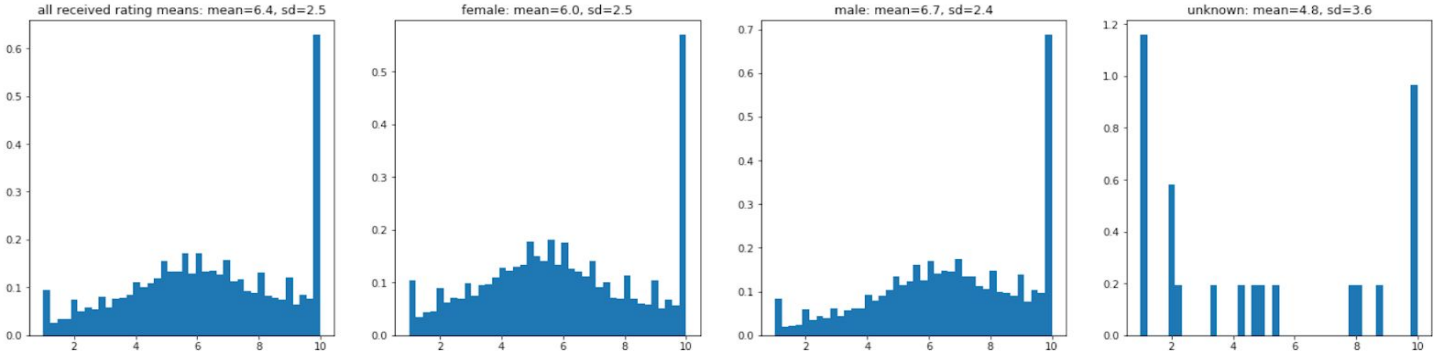
(~5%). However, without the true labels, we cannot determine the accuracy of our guess. But we kept the subset as another extra group to see if they inhabited some other interesting properties.

Rating distributions

The second analysis we did was into the ratings themselves. As we had previously grouped the ratings into a single mean value for both the given and received ratings, it did manifest excellent statistical properties as Central Limit Theorem would dictate.



As you can see from the rating distributions, they quite perfectly follow the standard distribution as you would expect. This is where the high node connectivity is helpful, as the large sample size leads to this type of convergence around the mean. Also, the mean values are slightly different for females and males (6.6 and 5.5) which you can take as men being bit stingier about their ratings in general..

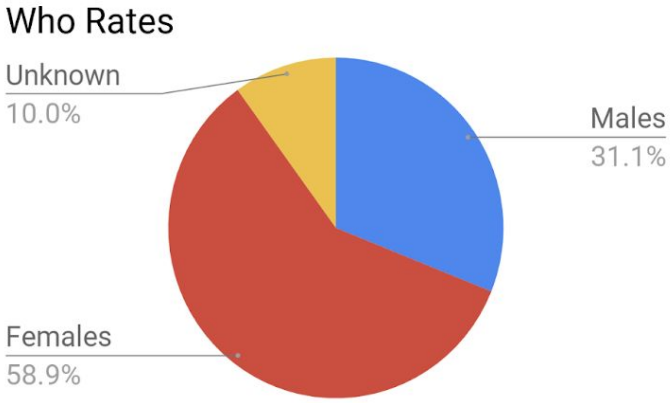


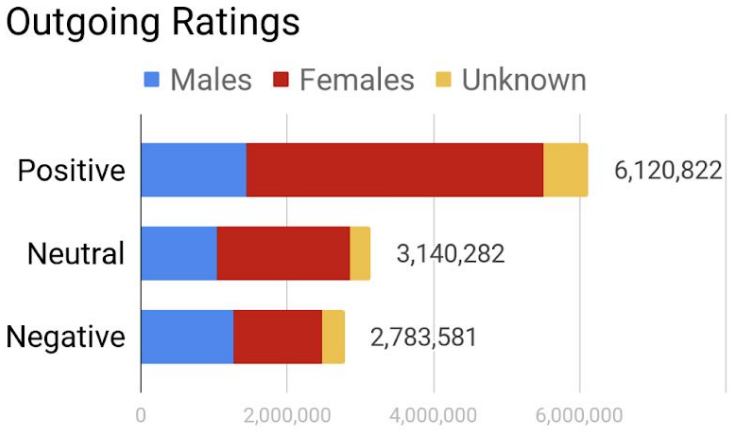
For the mean values of the received rating weights the situation is a bit different. There are a lot of nodes with small in-degree (53 % of nodes has in-degree less or equal to 5), which is why their means do not exhibit as strong convergence as the given ratings. A particular point of interest is the giant bin of 10 ratings - users with perfect 10 as their mean of received ratings. Intuitively it seems quite absurd that so many users are “perfect

tens", but it does seem that few users are consistently rated as 10 by all the other users.

Rating types

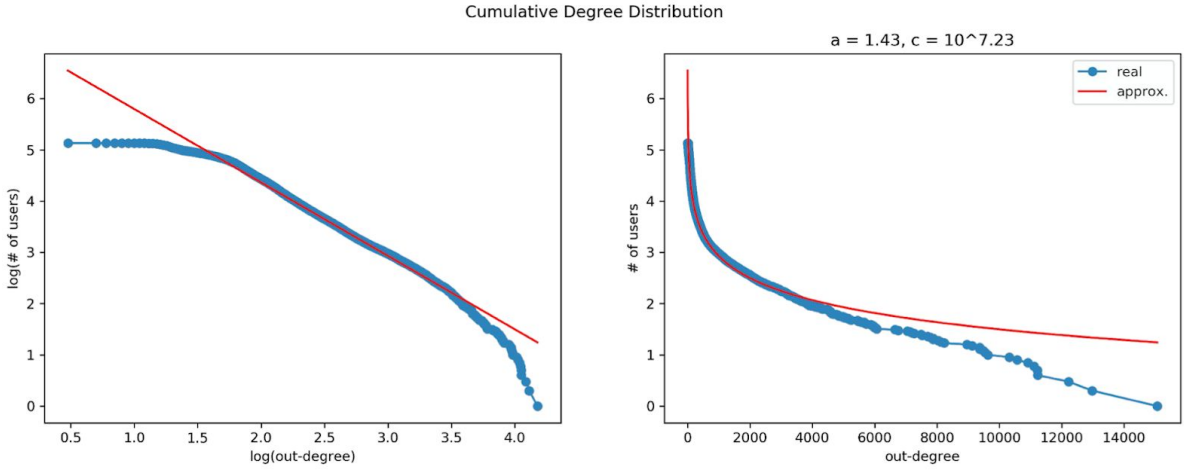
Another interesting question is what is the distribution of negative (weight 1-3), neutral (weight 4-6) and positive (weight 7-10) ratings. From the graph below we can see, that there are twice as many positive ratings as negative. The majority (67 %) of those positive ratings are issued by females and of all the ratings females give almost 60 %.





Degree distributions

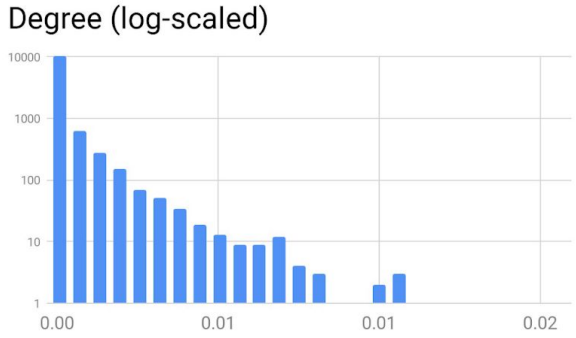
Thirdly we analyzed the out-degree and in-degree distributions of the nodes. As mentioned in the course lectures, these types of social graphs commonly follow a power-law distribution and ours was no exception.

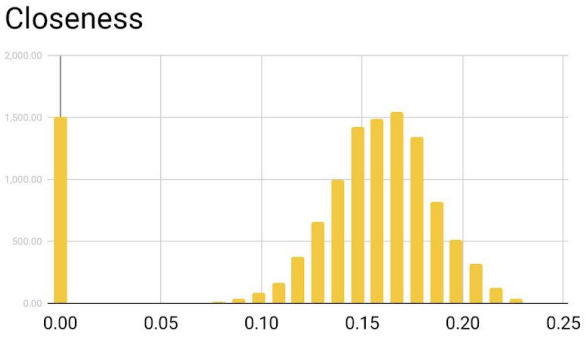


We estimated the parameters using the density function $P(k) = ck^{-a}$ to be (a = 1.43, $c = 10^{7.23}$) with linear regression applied on the log-log input (the left plot). The model fits the cumulative distribution of degrees quite well, although the curve's tail is slightly off. For social networks a is commonly between 2 and 3, here the value is 1.43 (descent of the curve is gradual).

Community analysis

Because our graph was too big, we had to divide it into smaller pieces for many of the graph analysis algorithms. Also, because the graph was directed, some of the algorithms were impossible to compute. For this analysis, we used the previous group of gay males as undirected graph.





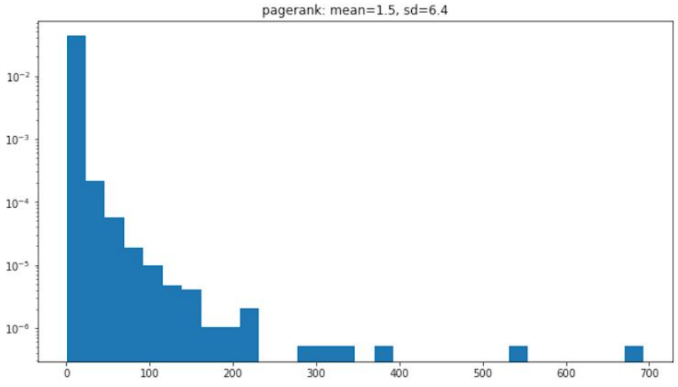
We computed the degree centrality and closeness centrality with quite pleasing results. It seems that unlike the lecture nodes would suggest, our distributions followed two completely different distributions that were not correlated.

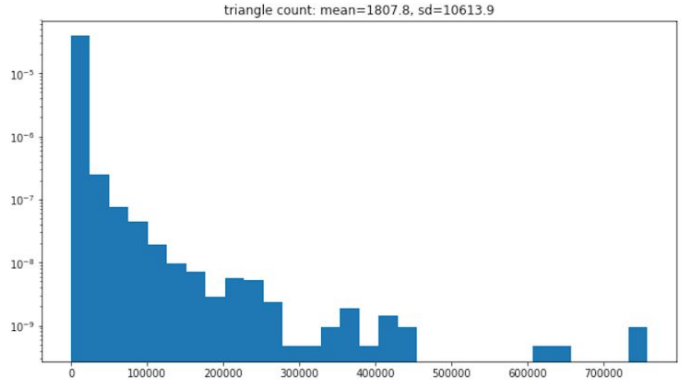
Other analysis

In the duration of this project, we did try many different approaches with mixed results. As we said earlier, our large graph was a major hindrance in our analysis. That caused us to waste time with nothing to show for it. Especially NetworkX performed poorly with even a moderately sized graph and took a long time to run.

On the other hand, we had great results with Graphlab, which ran its algorithms very quickly compared to NetworkX. For example, it was impossible to run Pagerank on the full dataset with NetworkX, yet with Graphlab it computed it in less than a minute. What was the reason for this, we do not know, but it was frustrating to notice a popular library so lagging in behind in their performance.

We computed both the Pagerank and triangle counts of the nodes, which pretty much followed the same power-law distribution as the node degree before.





We thought about clustering the nodes based on e.g. graph embeddings, but we did not have time to spare to work out our approach.

We ran a logistic regression model for determining the gender of the node based on the four features of indegree, outdegree, rating mean and received rating mean, which scored moderately at 0.75 against the original dataset. So you could infer the gender somewhat based on the features, but we wouldn't call the result in any way reliable.

Conclusion

All in all, it was an exciting project that taught us a lot about data analysis of graphs, and we feel satisfied with the results we got. Given more time, we could have maybe analyzed the graph properties even better and spent more time finding methods that would have worked with our large-sized graph.

Nevertheless, we managed to find some results to our original questions about if there were some differences between the different genders, and were the ratings evenly distributed. In short, the conclusion is this: if you are popular in a dating site, the chances are that you have a lot of incoming edges but not many outgoing. And that you are probably male if you are consistently rated 10 with many ratings. And females rate other people higher than males.