methods\_desc

### Methods description

As shown by …, the yule distribution provides a pretty reasonable fit to CD sales data.

The yule distribution arises from an underlying *snowball* process, which is defined as follows:

For each agent:  
 - With probability p, go to random artist, with probabilities proportional to previous playcounts for given artist. This is the "snowball" part.  
 - With probability 1-p, go to random artist, with equal probabilities. This part is for introducing randomness, as with p=1 first chosen artist would get all the listeners.

Fitting the yule distribution using MLE does not give any information about underlying p in this case. To obtain the estimate, we have simulated the process using 100 000 steps and 10 000 artists (same number as in the empirical data).

*Tutaj coś o tym dlaczego robimy to co robimy, do dopisania kiedy będziemy mieli całą podbudowę teoretyczną.*

Algorithm proposed by us is as follows:

1. For each agent randomly select a vertice of the graph (artist). Increment the playcount for each artist by number of agents who have chosen it.  
2. In next steps make each agent do one of the following actions:  
 - With probability p, go to random vertice, with probabilities proportional to previous playcounts for given vertice. This is the same as the "snowball" part in yule process.  
 - With probability 1-p, go to neighboring vertice (similar artist) to previously chosen one. If the artist from previous step does not have any neighboring vertice, go to random one (without weighting).  
3. For each vertice selected, increment the playcount for each artist by number of agents who have chosen it.

This simulation is designed to test if the recommendation system provided by last.fm site is contribution to some artists being exceptionally popular. An intuition behind this is that popular artists are more often recommended as similar to inspected one, as the popularity bias exists.

Results of the simulations were then compared to empirical distribution of playcounts. Distributions were compared using quantile-quantile plots, empirical distribution being the base one. As a reference, two theoretical distributions were fitted using maximum-likelihood estimation. These were Yule distribution and Pareto distribution. Both of them are referenced by … for obtaining reasonable estimation of stardom distribution.

More formal tests of goodnness of fit to the dataset were also provided. A widely popular choice for comparing two arbitrary distributions is a Kolmorogov-Smirnof test. However, for this particular dataset important information is contained in the upper tail. For such distributions, Anderson-Darling test is more sufficient, and thus was also used.

R packages igraph, tidyverse, fitdistrplus, gamlss.dist were used.

### Results

In all simulations using graph structure, constant number of agents and number of steps were used, set at 10 000 and 50, respectively. After 50 steps the distribution of playcounts was not changing substantially anymore. Probability of selecting random popular artist *p* was initially tested for p = 0.1, 0.2, …, 1. After this procedure, quantile-quantile plot analysis have shown that reasonable p < 0.1.

Value of parameter *mu* describing Yule distribution obtained through maximum likelihood estimation was 1898.96.

Table containing parameters of K-S test and Anderson-Darling are shown in the table.

As it can be seen from the qq-plot, the simulation results resemble empirical results closely. The difference in analyzed distributions lies in the tail - obtained playcounts differ from the empirical data.

Another interesting result is comparison with snowball process - probability of particular agents “attaching” to the main trend needed to reflect empirical distribution is at 0.95. For random graph walk it is only 0.07 . This means that even without strong drive for listening to popular music, there exists a mechanism (recommendation system), for which the distribution of popularity can be also obtained.