Attachment Mechanisms in Catalogue Networks

Matthew Lowe

September 2009

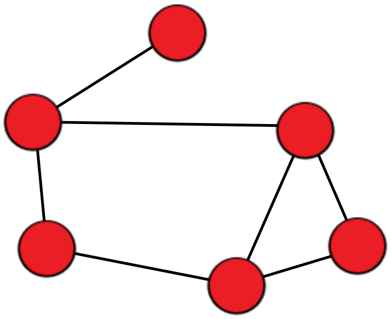


Figure 1: An example of a network with 5 nodes and 7 edges

Introduction

A network is a configuration of agents and the relationships between the agents. Agents are represented by *nodes* or *vertices* with a relationship between two objects represented by an *edge* linking the two. Two nodes are said to be connected if an edge exists linking the nodes. Edges can be directed, meaning they represent a relationship in one direction only between nodes, or can be weighted, meaning the edge has a value, or *weight*, associated with it. In a network with weighted edges the weight of the edge is considered rather than just whether it exists. The *degree* of a node in an unweighted network is the number of edges with one end attached to the node. Figure 1 shows an example of a basic undirected network, other examples of networks include the world wide web, social networks and food webs.

Often a network cannot be considered to be fixed, and it's evolution and growth needs to be considered. One important tool for doing this is *preferential attachment*, the first of application of this in a growing network is thought to have been Price[1] in 1976 although it has also been considered previously by Simon[2] in 1955 and Yule[3] in 1925. The term *preferential attachement* was coined by Barabasi and Albert[4] in their 1999 paper on the growth of the world wide web. In a simple application of preferential attachment, a new node with one edge is added to the network, connected to an existing node with probabilities proportional to the degree of the node. For a network with N nodes the probability of choosing node n_i with degree k_i is

$$\pi(n_i) = \frac{k_i}{\sum_{j=1}^{N} k_j}$$

This mechanism means the higher the degree of a node, the more new edges the node will attract. As well as adding new nodes to a network this can also be used to add new edges to a network with a fixed number of nodes.

A further generalisation of networks discussed so far is a *bipartite network*. A bipartite network is a network with two distinct sets of nodes, where edges can only exist between nodes from different sets (e.g. Figure 2).

Netflix Data

In October 2006, the online DVD rental website Netflix announced a $1 million competition to try and improve it's recommendation algorithm by 10%. A user of Netflix has the option of rating any DVD an integer number of stars, with 1 being the lowest rating and 5 being the highest rating. As part of the Netflix

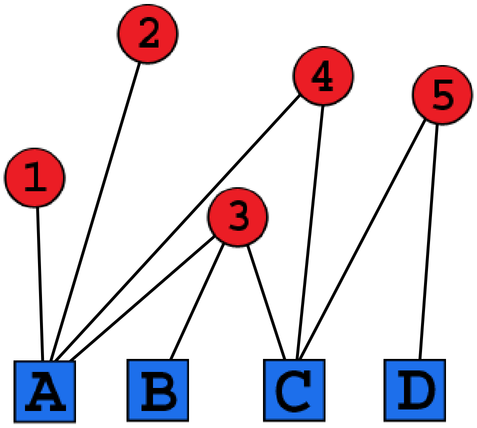


Figure 2: An example of a bipartite network with node sets $\{A, B, C, D\}$ and $\{1, 2, 3, 4, 5\}$

competition they released a data set revealing how 480,189 users rated up to 17,770 DVDs, this consisted of 100,480,507 ratings dated from November 1999 until December 2005. Each piece of data consists of a user ID, which movie they rated, the date they rated it and how many stars they gave the movie. Throughout this paper, the term 'movie' or 'film' will refer to any of the 17,770 DVDs in the Netflix catalogue, although these also include TV series, music videos, documentaries etc.

A Catalogue Network Model

This data can be represented by a bipartite network, with the two node sets being the set of users, $\mathcal{U} = \{u_1, ..., u_U\}$ and the set of movies $\mathcal{M} = \{m_1, ..., m_M\}$, where M is the number of movies in the set and U is the number of users. This can be considered either as an unweighted network, with a connection between a movie and a user if that user has rated the movie and no connection if not, or as a weighted network where the number of stars a user has given a movie is the weighting on each edge.

In order to better understand this network it is important to consider the dynamics and evolution of the network. In 2009 Mariano Diaz[5]proposed a dynamic model for wiring and attachment in a bipartite catalogue network to represent the Netflix data. The model starts with a fixed set of users and movies initially disconnected, edges are then added one at a time between unconnected nodes according to a set of rules until a chosen time limit or until all the nodes are connected. Adding an edge between a user and movie represents a user giving a rating to that movie.

A combination of uniform random attachment and linear preferential attachment was proposed to select the nodes. When chosing the movie nodes, preferential attachment represented social popularity of movies, movies that are watched by many are more likely to be socially popular and therefore attract more people to watch, and subsequently rate. Uniform random attachment was a representation of users personal tastes, movies a user wants to watch regardless of what other people choose.

The user is chosen using preferential attachment with probability q, and the movie is chosen using preferential attachment with probability p, where $p, q \in [0, 1]$. So at each time step a user of degree h_i is chosen with probability

$$P_{\mathcal{U}}(t, h_i) = \frac{1-q}{\widehat{U}(t)} + \frac{q \cdot h_i}{\sum\limits_{u_j \in \widehat{\mathcal{U}}} h_j(t)} \tag{1}$$

Where
$\widehat{\mathcal{U}}(t) := \{\ u \in \mathcal{U} \mid \textit{degree of } u \ \neq \ M \textit{ at time } t\ \}$ and $\widehat{U}(t) = ||\widehat{\mathcal{U}}(t)||$. And the movie of degree k_i is chosen with probability

$$P_{\mathcal{M}}(t, k_i) = \frac{1-p}{\widehat{M}(t)} + \frac{p \cdot k_i}{\sum\limits_{m_j \in \widehat{\mathcal{M}}} k_j(t)} \tag{2}$$

Where
$\widehat{\mathcal{M}}(t) := \{\ m \in \mathcal{M} \mid \textit{degree of } m \ \neq \ U \textit{ at time } t\ \}$ and $\widehat{M}(t) = ||\widehat{\mathcal{M}}(t)||$. $\widehat{\mathcal{U}}$ and $\widehat{\mathcal{M}}$ are used instead of $\mathcal{U}$ and $\mathcal{M}$ to prevent nodes which are fully connected from receiving new edges.

Movie Ratings

An element missing from this model is an attachment mechanism driven by the movie's ratings. Each movie in the Netflix dataset has received a number of ratings from users. The mean number of ratings is around 5600, the minimum number of ratings is 3 and the movie rated most often has 232,944 ratings. For each film, taking the mean value of these ratings will give the film an intrinsic value which can be used to drive an attachment mechanism. The term 'mean rating' or 'fixed rating' shall be used to describe this value. When modelling it will be necessary to give each movie a fixed rating, representing the movie having already been rated by a large number of users and the mean of those ratings have been taken. Figure 3 shows that the distribution of mean ratings fits a Weibull distribution well. Thus each movie in $\mathcal{M}$ can be given a fixed rating drawn at random from the Weibull distribution.

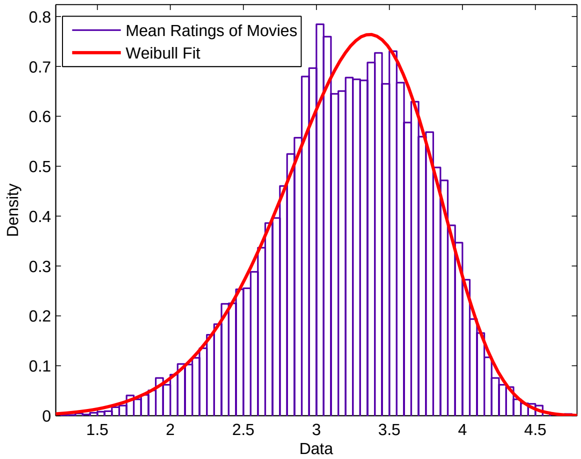


Figure 3: Distribution of the mean rating of 17770 movies given by 480,189 users, fitted by a Weibull distribution with *shape parameter* $k = 7.07$ and *scale parameter* $\lambda = 3.44$

Attachment driven by ratings

For each movie in the set of movies, $m_i \in \mathcal{M}$, an attachment mechanism must assign to each movie a value, ρ_i, such that

$$\hat{\rho_i} := \frac{\rho_i}{\sum\limits_{m_j \in \mathcal{M}} \rho_j}$$

is the probability of movie m_i being chosen. To create an attachment mechanism driven by ratings we shall use $\rho_i = \rho(r_i)$, where r_i is the movie's fixed rating. Determining what is the most suitable form for ρ is not a straight forward question, it asks what a person judges a rating to be worth relative to other ratings. For instance if ρ is a linear function of r then this represents a person judging a 4 star film to be twice as worthwhile to watch compared with a 2 star film, as the four star film would have twice the probability of being chosen as the two star film. How much worth a user considers a rating to have would almost certainly change from person to person - but we shall not be considering these potential variations for now and instead look at what happens on average for a large number of people.

Figure 4 shows how a movie's mean rating varies with the number of ratings it has received. The left hand side shows the raw data, and the right hand side takes an average of the data. It is clear from the graph that on average films with a higher rating get more ratings. The best fitting curve to the left hand side of Figure 4 was a curve proportional to $e^{r^{1.2}}$. Therefore $\rho(r_i) = \exp(r_i^{1.2})$ is used to generate the attractiveness of a movie's mean rating and thus the probability of a movie being chosen, when using attachment driven by ratings.

A Network Model with Ratings

Using a combination of preferential attachment and attachment driven by ratings, a new bipartite catalogue network model could be created. Users are still chosen through a combination of uniform random

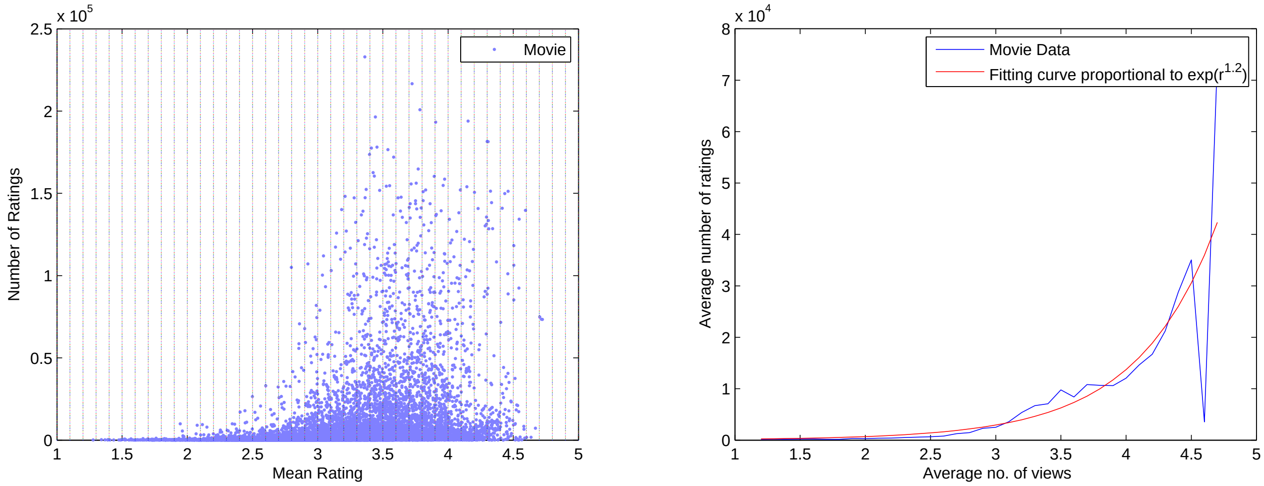


Figure 4: The left hand graph shows a plot of mean rating v number of rating for all 17770 movies. Taking the average number of ratings for each movie in the vertical columns produces the right hand graph, the right hand graph shows a curve fit $\propto e^{r^{1.2}}$

and preferential attachment, preferential attachment is used with probability q, where $q \in [0,1]$. Movies are now chosen with probability p of preferential attachment and probability $1-p$ of attachment driven by ratings, where $p \in [0,1]$.

From an initially unconnected network with $\mathcal{M}$ the set of movies and $\mathcal{U}$ the set of users, edges are added one at a time. At a each time step, t, a user with degree h_i is first chosen with probability

$$P_{\mathcal{U}}(t, h_i) = \frac{1-q}{\widehat{U}(t)} + \frac{q \cdot h_i}{\sum\limits_{u_j \in \widehat{\mathcal{U}}} h_j(t)} \qquad (3)$$

where, again,
$\widehat{\mathcal{U}}(t) := \{\ u \in \mathcal{U} \mid \textit{degree of } u \ \neq \ M \textit{ at time } t\ \}$ and $\widehat{U}(t) = ||\widehat{\mathcal{U}}(t)||$. Once a user, u_i, is chosen then a movie, with degree k_i and fixed rating r_i, is chosen with probability

$$P_{\mathcal{M}}(t, k_i, r_i) = \frac{p \cdot k_i}{\sum\limits_{m_j \in \widehat{\mathcal{M}}} k_j(t)} + \frac{(1-p) \cdot exp(r_i^{1.2})}{\sum\limits_{m_j \in \widehat{\mathcal{M}}} exp(r_j^{1.2})} \quad (4)$$

This time
$\widehat{\mathcal{M}}(t) := \{\ m \in \mathcal{M}\ |\textit{degree of } m \ \neq \ U \textit{ and m is not already connected to } u_i \ (\textit{at time } t)\}$

Once more preferential attachment represents what is socially popular, and attachment driven by ratings represents users picking a movie because it is considered a 'good' movie and is generally recommended. As attachment driven by ratings still allows badly rated films to be chosen, just with smaller probabilities, this does provide an element of personal taste independent of what other people think. It still allows a user to choose a 2 star film over a 4 star film, but it is just rarer.

Results

To see how this model compares to the dynamics of the actual Netflix data, samples had to be taken from the Netflix dataset. 500 samples were drawn from the data each of which contained 10 movies and 100 users. The *'birth'* of a movie or user shall be defined to be when the movie or user first appears in the Netflix data set, so that a user born on 1st January 2003 means the user rated their first movie on the 1st

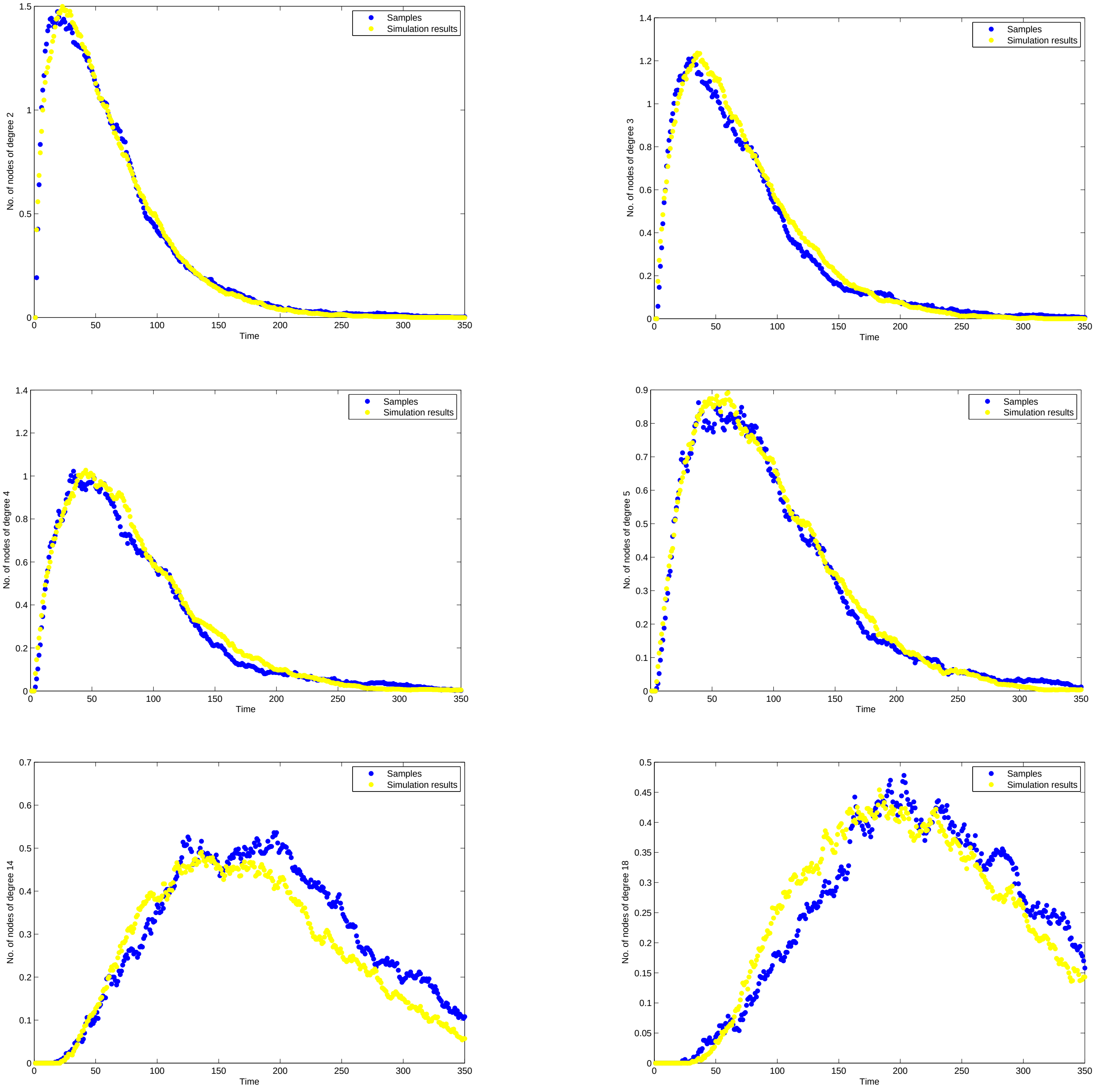


Figure 5: Movie degree distributions of the sampled data compared with numerical simulations. Simulations were run with $M = 10$, $U = 100$ and optimal parameters $q = 0.74$ and $p = 0.53$ were found. Simulations show a good fit to the data for low degrees, with less accuracy for higher degrees. From top left to bottom right; degree distribution with degree = 2,3,4,5,14 and 18

January 2003.

To generate each sample, 10 movies were chosen at random from all movies born before 1st Jan 2003 and which have more than 10,000 ratings. The set of ratings of these movies were then restricted to the ratings that occurred after Jan 1st 2003, and then further restricted to ratings by users born before Jan 1st 2003. From the set of users who made these ratings, 100 users were chosen at random and only the ratings they made of the 10 movies considered. This data can then by represented by a bipartite network with fixed catalogues of users and movies, where edges are added as users rated movies.

These 500 samples contained between 175 and 670 ratings each, with a mean of 400 so a fully connected network was never achieved. Where each time step, t, represents one edge being added, the number of movie nodes with degree k at time t is denoted by $N(t, k)$, this is known as the movie *degree distribution.*

Simulations were then run using $M = 10$ and $U = 100$, and the average degree distribution from the simulations were compared with the average degree distributions from the samples. Figure 5 shows the results of these simulation with $q = 0.73$ and $p = 0.54$, which were calculated to be the best fit.

Further Study

New Release Effect

An element this model does not capture is the effect of a new movie entering the catalogue. Figure 6(a) shows a typical example of a new feature film[1] entering the Netflix catalogue. It shows an initial peak in popularity which gradually declines to a steady rate. This shape can be seen for the majority of feature films, and suggests there is an exogenous effect which can be due to advertising both from the films distributors and the film appearing more prominently on the Netflix website as a 'New Release'.

If the model were to include a mechanism so that $\mathcal{M}$ was no longer of fixed size but increasing, then a 'false shock' should be added to account for the generally higher popularity of movies on their initial release.

For television series, an initial high popularity is also often seen, however as figure 6(b) demonstrates, a two or more peaked graph is more widely seen. These extra peaks are also likely to be from exogenous effects that the model does not capture, a new series or rerun of the show on television or a new series release on DVD are all likely to cause renewed popularity in a television series. Figure 6(c) shows how the release of a new series of Gilmore Girls renewed the popularity of an older series, this is another important exogenous effect.

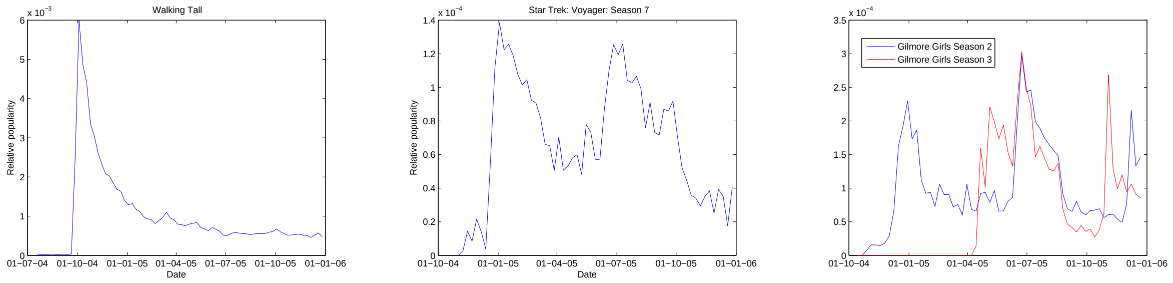
Preferential attachment with memory

Preferential attachment in this model is a representation of social popularity or social trendiness. This is something that is always changing, the social status of movies is always changing, with new blockbusters coming in to replace old ones as the most talked about movie. For this reason, it is worth considering the use of preferential attachment with memory.

Preferential attachment mechanisms like this have been suggested before[6] when studying academic paper citation networks. Two methods were proposed, the Gradually-vanishing Memory Preferential Attachment Mechanism (GMPAM) gave weights to citations based on how recent they were, and the Short-term Memory Preferential Attachment Mechanism (SMPAM) used only citations within the last year within a preferential attachment mechanism.

Using a preferential attachment with memory could be particularly useful when combined with initial 'false shocks' given to new releases. Simulations were run using the model created but with an initial false memory to give a movie an intially high popularity as if it were just released, and with SMPAM replacing the standard preferential attachment. As expected the longer the memory was in the SMPAM the longer the film's popularity took to decline. Different rates of decline from a movie's initially high popularity are also seen in the data. Further study would be needed to determined whether this is linked to other properties of the

[1] as opposed to a television series or other DVDs within the Netflix dataset



(a) Time series popularity of typical feature film

(b) Time series popularity of a typical television series

(c) The release of a new series causes an increased popularity in older series

Figure 6: Relative popularity is measured by *Number of ratings of movie* divided by *Number of ratings of all movies* in each week.

movie such as it's rating, but if it is found to be an independent factor using different length memories for different movies could be a way to capture this effect. GMPAM is also worth considering, although it is more complex it makes more sense for a movie to gradually fade from societies popularity 'memory' rather than just disappear after a certain of time

User habits

In order to better understand how a user will choose which film to rate it may be useful to take a user based perspective on the data. Different users have very different habits when it comes to rating a movie, for instance Figure 7 shows examples of how the distribution of different users' ratings vary. Therefore ideally any modelling of this data should not treat users in the same way. First it is worth considering whether these differences in rating distributions are caused by fundamental differences in different users criterion for rating films, or can it be said that people who give lower ratings are just watching (and hence rating) worse films than those who are giving higher ratings? One way to investigate this is to measure how much each user disagrees with the consensus on how good a film is. For a particular user, their '*adjustment*' will be the average of *the rating they have given a movie* minus *the mean rating of that movie* for every movie they have rated. A large positive *adjustment* means the user is overrating films compared with the average user, a large negative *adjustment* means the user is generally underrating movies. A *adjusment* close to zero shows that a user is neither consistently giving high or low ratings.

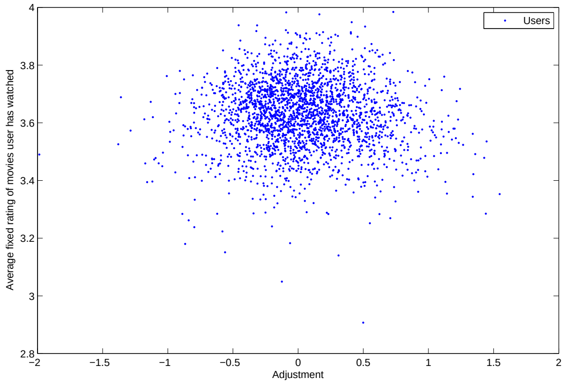


Figure 8: For 2452 users, their adjustment is plotted against the average of the mean ratings of each film they have watched. The y-axis is an indication of the quality of movies a user tends to watch. Each user in this sample has rated over 100 movies.

The relative symmetry through the line *adjustment* = 0 in Figure 8 suggest that there is no reason to believe people who consistently rate badly are more likely to be watching worse films than those rating consistently high, in fact it would

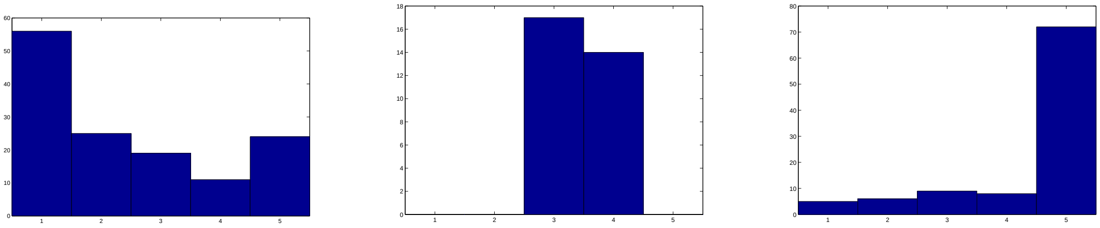


Figure 7: Distribution of ratings for different individual users

appear that what quality of films a user watches is largely independent of how highly they are then likely to rate a film. The graph also shows that there are relatively few users who are over-raters or under-raters with the vast majority having an adjustment less than 0.5 from zero. Along with the small range of the y-axis this suggests a randomly selected user is likely to have a rating pattern not too far off the 'average user'.

Ratings changing with time

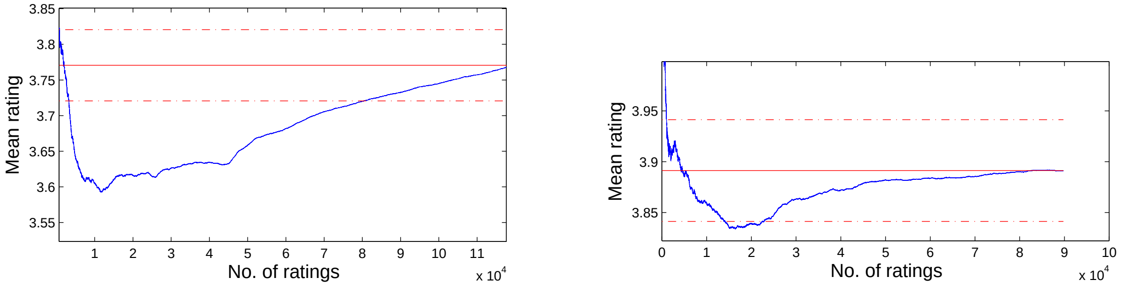
As mentioned earlier, giving a movie a fixed rating to represent it's mean rating at the beginning of a simulation is not true to reality. Every time a user rates a film, it will slightly change what the mean rating is. Figure 9(a) shows how much a movie's mean rating can change even after it has received a large number of ratings. In this example the movie's mean rating after 40000 viewings is 0.15 star below it's final mean rating. Whilst 0.15 stars may not be a huge change, in the model this difference would be greatly amplified when the attractiveness of the movie is determined. This indicates that taking the movie's mean rating to be fixed in the model might introduce flaws into the results. Figure 9(b) shows an example of a movie which has a more stable mean rating. After just a few dozen ratings the mean rating barely goes outside of a 0.05 interval either side of the final mean rating. Both movies display the same shape of a mean rating that initially dips and then grows, this pattern is surprisingly common amongst movies. This could be worth investigating further.

Conclusion

The aim of this project was to introduce a mechanism for including the values of the ratings of movies into an attachment mechanism in a network catalogue model. This was achieved, with results indicating this was an important factor to consider. The model fitted the actual data best with preferential attachment used 53% of the time and an attachment mechanism driven by ratings 47% of the time.

Further factors that might prove important were also investigated although without inclusion into the model. The effect of a new movie being introduced was discussed, it suggested that having a catalogue model which expanded was not a simple extension as many new movies had a large peak of popularity upon their introduction into the netflix catalogue. In order to capture this in the model a way of investigating and measuring these shocks would need to be developed.

Another factor that was considered was having a preferential system with memory to reflect changing popularities in movies. When combined with the idea of movies having popularity shocks upon their release, a longer memory meant the movie's popularity took longer to decline. Some movies taking longer to decline in popularity suggests they are bettered remembered, whether this feature of different rates in popularity decline are related to other factors or whether it is an independent feature would need further investigation, but if it were found to be an independent feature the use of different short term or gradually fading memories could be a way of reflecting this feature in a model.



(a) How the mean rating of 'Enemy of the State' changes.

(b) How the mean rating of 'Chocolat' changes.

Figure 9: How the mean rating of movies change with the number of ratings they received. The solid line is the mean rating after all ratings from the data set, with the dashed lines indicating an interval of 0.05 either side of the mean

User habits were also looked at, a user based perspective was hoped to provide a different insight into the data. It was concluded that looking at how a user rates each film is not a useful way to predict which movies they are then likely to rate in the future, and further insights into user habits would be needed to produce a model which treats different users in different ways. However it was also found that most users do not behave substantially differently to each other.

In the model each movie was given a mean rating that was fixed, in reality the mean rating changes with each new user rating it. Looking at this further suggested that in many cases the mean rating changed sufficiently significantly with further ratings that this could be a factor worth including in a more complex model.

Acknowledgements

I would like to thank Mason Porter and Nick Jones for their input, help and support in this project, Mariano Diaz for his data samples, coding, time and input; and Noel Cody for his thoughts and ideas.

References

[1] D. J. d. S. Price, Science 149, 510 (1965).

[2] H. A. Simon, Biometrika 42, 425 (1955).

[3] G. U. Yule. "A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S.". Philosophical Transactions of the Royal Society of London, Ser. B 213: 2187 (1925).

[4] A.-L. Barabasi and R. Albert, Science 286, 509 (1999).

[5] M Diaz, M Porter and JP Onnela. Competition for Popularity in Catalog Networks (2009). arXiv:0906.4675.

[6] M Wang, G Yua and D Yua. Measuring the preferential attachment mechanism in citation networks (2008). doi:10.1016/j.physa.2008.03.017