Modeling and Inference of Networks of the Internet Movie Database with Preferential Attachment Model

Fengnan Gao*

Mathematical Institute
Leiden University
The Netherlands

We study the movie-actor network with the data from the Internet Movie Database. The movie-actor network is an example of *collaboration networks* and thus studying it provides us insights into the organization of such networks. We are interested in proposing a suitable model to describe the dynamics of the movie-actor network, how the network comes to what it is today. Through empirical study of the dataset, we come to a modify preferential attachment model to accommodate behaviors of the dataset and propose a general model for collaboration networks. We show that our model captures more-or-less the power-law degree distribution that is observed in the real dataset. Simulations and theoretical studies are present to verify that our model is satisfactory.

1 Introduction

A network is a structure containing vertices and edges. A complex network, unlike simple networks, is a graph/network with non-trivial topological features, such as *scale-free* ([1]) and *small-world* ([13]). Empirical studies on real networks, such as the Internet, the World-Wide Web, social and sexual networks, and networks describing protein interactions, show fascinating similarities: in particular, many of these networks are both *small-world* and *scale-free*.

*gaof@math.leidenuniv.nl

The collaboration network, is a category of networks, where the vertices are people and the edge among two vertices stands for a "*collaborative*" relation – whether they have "collaborated". A perfect example is scientific collaboration networks – see [9, 8, 10, 11] – where scientists participate as vertices and the scientific acquaintances, typically whether they have *collaborated* in writing the same papers, are the edges. We consider another form of collaboration networks – the movie-actor networks: the actors and actresses are vertices and the "acquainted" relation, whether they have played in the same movie together, play as edges. It has been studied by several papers, most notably in [3], where the scale-free property of the movie-actor network has been investigated.

In this paper we study the well-known Internet Movie Database (IMDb) to find out more the movie-actor network. The IMDb collects all the information relating to movies (consequently actors) and formulate the data in a way in which the public, including us, can make use of it. It is also worth noting that results in this paper are reproducible for everyone due to the public availability of the dataset. The files, which one can download from the website of the IMDb or its ftp mirrors, are plain text files with each line representing one actor or one movie, or the appearance of one actor in one movie. We write Python scripts to parse the lines and store the extracted information into a well-structured database. Our primary interest is the topological structure of actor network where each actor is vertices and each edge is the collaborated relationship. Hence the aforementioned structure is designed to facilitate the accessing of interactions between actors. We utilize the same structure to store information for simulations.

The degree distribution of some networks are often modeled by a power-law, which means the probability p_k of a certain degree k appearing is of the form $p_k \approx Ck^{-\tau}$ for k sufficiently large and C and τ being both positive constants. It has been observed in a lot of networks, in social, biological and information networks (references go here). It is easy to see $\log p_k = \log C - \tau \log k$, i.e. $\log p_k$ decreases with respect to $\log k$ with slope $-\tau$. Consequently whether a degree sequence follows a power-law is often determined by loglog-plot of degree frequency versus degree. The power-law exponent τ is often obtained by performing a linear regression on data of $\log p_k$ against $\log k$. There exist in the dataset of the IMDb some power-law-like degree distributions of essential importance to our modeling. A possible underlying model that might be responsible is preferential attachment model (PAM). In PAM, new vertex is linked to existing vertices with a probability proportional to their degrees l, i.e. $p(l) = \frac{l+\delta}{\sum_j (d_j+\delta)}$, where d_j is the degree of vertex j and $\delta \geq 0$ is to give the model more flexibility. The model was first proposed by [3] and has been studied extensively so far. It was an attempt to define a mathematical model characterizing so-called "Mathew effect" where "the rich get richer".

To model the IMDb, the fundamental approach is to think of the network as a graph of vertices participating in the network in a somewhat random manner. Hence we seek a random graph model that describes the dynamics of collaboration graphs, is simple (i.e., has only few parameters) and is sufficiently flexible. We propose a dynamic model modified from the PAM, which to some extent explains the evolution

of the movie-actor network.

The paper is organized in the following way. In Section 2 we present the novel conceptual model. While studying the dataset offered by the IMDb, we fit the model to our dataset of the IMDb in Section 3. Simulations of the model and comparisons between the simulation and real dataset validates, up to a certain degree, the fitness of our model. This conclusion is further consolidated in Section 5 where we prove, with proofs in a separate attachment, that the model indeed leads to a power-law degree distribution.

2 *Conceptual Model* Description

In this section we describe a model in a skeleton-like fashion. The general model without giving the exact specification is hereafter referred as the *"conceptual model"*.

We denote t as an index for the movie network measuring the growth of the network, i.e. t is the number of movies but is *not* real calendar time. We consider a comprehensive movie network $(\mathfrak{G}_t)_{t\geq 0}$ ($\mathfrak{G}_t = \left(\mathcal{M}_t, \mathcal{A}_t, \mathcal{E}_t^{\mathcal{M}}, \mathcal{E}_t^{\mathcal{A}}\right)$) with two layers:

1. The layer of movies $\mathcal{M}_t$, with vertices $(m_\alpha)_{\alpha\in\{1,2,\ldots,t\}}$ as movies.
2. The layer of actors $\mathcal{A}_t$, with vertices $(a_\beta)_{\beta\in\{1,2,\ldots,\phi_t\}}$ as actors, where $\phi_t = |\mathcal{A}_t|$ is the size of actor network at time t.
3. The edge set $\mathcal{E}_t^{\mathcal{M}}$, describing the relation between movies and actors.
4. The edge set $\mathcal{E}_t^{\mathcal{A}}$, describing the relation among actors.

The edges in the layers are as follows:

- Movie m and actor a are linked if and only if actor a has played a role in movie m, which we denote as $a \leftrightarrow m$, and this link $a \leftrightarrow m$ defines an edge $e_{a\leftrightarrow m} \in \mathcal{E}_t^{\mathcal{M}}$ between a and m.
- Two actors $a_1 \in \mathcal{A}_t$ and $a_2 \in \mathcal{A}_t$ ($a_1 \neq a_2$) are neighbors if and only if there exists a movie $m \in \mathcal{M}_t$ such that a_1 and a_2 are both linked to m. We write $a_1 \leftrightarrow a_2$ to denote that actors a_1 and a_2 are neighbors and this link defines an edge $e_{a_1\leftrightarrow a_2} \in \mathcal{E}_t^{\mathcal{A}}$.

We define the appearance time $T_M(m)$ of movie m as $T_M(m) = \min_{m\in\mathcal{M}_t} t.$ and the movie size $S(m)$ as the number of actors playing in it $S(m) = \sum_{a\in\mathcal{A}_{T_M(m)+1}} \mathbf{1}_{\{a\leftrightarrow m\}}$. We define the appearance time $T_A(v)$ of actor v as $T_A(v) = \min_{v\in\mathcal{A}_t} t$.

The actor degree $D_t^A(v)$ of actor $v \in A$ at time t is defined as $D_t^A(v) = \sum_{a\in\mathcal{A}_t, a\neq v} \mathbf{1}_{\{a\leftrightarrow v\}}$, whereas the movie degree $D_t^M(v)$ of actor $v \in A$ at time t is defined as $D_t^M(v) = \sum_{m\in\mathcal{M}_t} \mathbf{1}_{\{v\leftrightarrow m\}}$. By convention, both degrees are defined as 0 when $t < T_A(v)$ (i.e. the actor has not appeared in the network).

The evolution starts from $t = 1$ with one movie and several actors in $\mathcal{M}_1$, and $\mathcal{M}_t$ is updated in the following way when a new movie m_{t+1} is introduced:

1. The number of actors ζ_{t+1} of the movie m_{t+1} (i.e. $\zeta_{t+1} = S(m_{t+1})$) follows a certain distribution ζ, independent from the past state $\mathfrak{G}_t$.

2. The number of new actors ξ_{t+1} (here we abuse the word "*new*" and "*old*": an *old* actor means an actor who has been in the network before this movie and a *new* actor means the contrary) of the movie m_{t+1} (i.e. the number of new vertices introduced to the actor layer by movie m_{t+1}) is a random variable, independent from the past $\mathfrak{G}_t$, and follows a distribution depending only on ζ_{t+1}.

3. (PAM) For $\psi_{t+1} = \zeta_{t+1} - \xi_{t+1}$ actors, who must be chosen from the actor layer $\mathcal{A}_t$. The *old* actors are chosen from $\mathcal{A}_t$ independently and by using the preference that depends on a certain quantity of the actors

$$\mathbb{P}(a \leftrightarrow m_{t+1} | \mathfrak{G}_t) \propto f(a), \tag{1}$$

   for any $a \in \mathcal{A}_t$, where f is a function depending on certain attributes of actor.

4. Add ξ_{t+1} new actors to the actor layer and adding the corresponding edges between actors and m_{t+1} and among the actor layer.

After movie m_{t+1}, $\mathfrak{G}_t$ evolves into $\mathfrak{G}_{t+1}$ with $|\mathcal{A}_{t+1}| = \phi_{t+1} = \phi_t + \xi_{t+1}$ actors and $|\mathcal{M}_{t+1}| = t + 1$ movies.

It is worth pointing out that the actor-layer of our double-layered network is a self-contained conventional collaboration network and actor degree is vertex degree in the collaboration network.

3 Empirical Fitting of the Model to the IMDb Dataset

We study the movie-actor network given by extracting the information out of the IMDb, and in due process we specify the distributions of ζ and ξ and the preference function f so as we have a concrete model.

3.1 Movie Sizes

First we study the movie size (the number of actors in one movie), i.e. the distribution of ζ in the conceptual model.

We see from Figure 1 which contains the movies sizes in one year that there is a peak at around 5 and after that the frequencies decrease with respect to movie sizes. It also seems that the distribution has a heavier tail than normal distribution. We do a more revealing loglog plot of p_k versus k with k being movie size and p_k being the frequency of movie size k.

From the look, it seems the movie sizes follow a power-law distribution because the tail looks like a straight line on the loglog plots when the number of movies goes up. We performs a regression on $\log k$-$\log p_k$ data evidenced in Figure 3 where the data of all movies up until year 2007 is taken into account. We see a straight line fits rather well in the middle of data range but suffers from a bad fit on the tail. We come to the

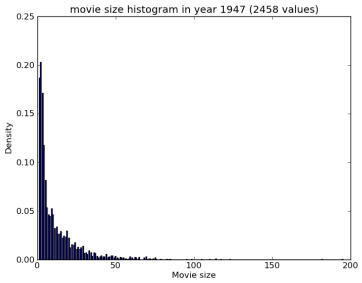
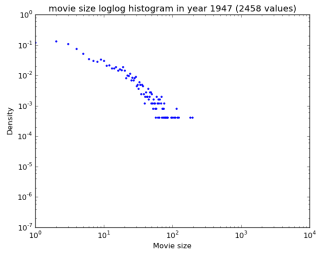
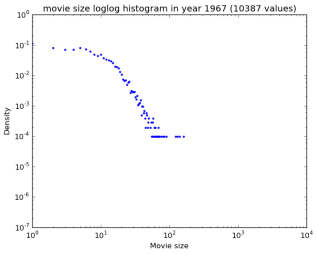


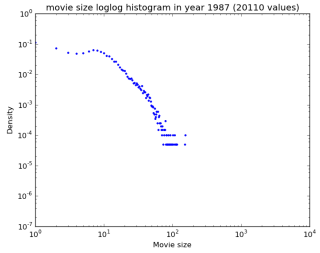
Figure 1: Histogram of movie sizes in 1947 with 2458 movies



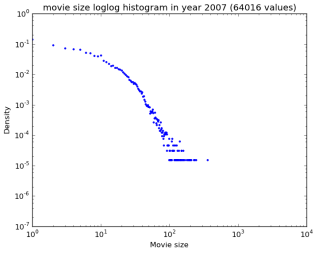
(a) 1947 with $2,458$ movies



(b) 1967 with $10,387$ movies



(c) 1987 with $20,110$ movies



(d) 2007 with $64,016$ movies

Figure 2: `loglog`-Histogram of movie sizes in 60 years from 1947 to 2007

conclusion that movie sizes do follow somewhat a power-law distribution but not an exact power-law.

For our purpose, we use the empirical distribution of all movie sizes as our ζ. It is interesting the movie sizes follow a power-law, but we are not interested in explaining it. Given that there are more than a million movies in the network, the empirical distribution should be close to the true distribution.

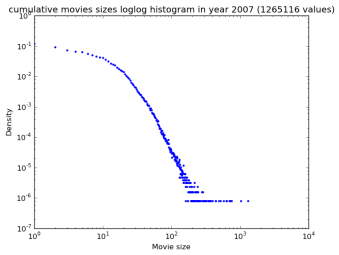


Figure 3: `loglog`-histogram of all movie sizes until the end of 2007

3.2 Number of new actors

We also need to try to find a good model for the generation of new actors, which are responsible for the expansion of the network. The question is that given a new movie has n actors, which is the distribution of number of new actors?

We give a scatter plot of the ratios r of the number of new actors to the movie size of all movies in year 1971 as in Figure 4. The ratio is not constant or close to a constant even when the movie size is large in any sense. In fact the ratio seems to have a lot of variations and spread around over the curves $xy = 1, 2, 3, 4 \cdots$ if we view the horizontal axis to be the x-axis and the vertical one to be the y-axis. However the latter observation is natural since both the movie size and number of new actors are integers. Therefore given $\zeta = n$ the movie size, ψ is not a binomial distribution but rather the ratio ψ/ζ distributes all over $[0, 1]$.

In light of the above observations, we propose a randomized binomial model. The idea is to put the ratio ψ/ζ at random in $[0, 1]$ then to perform a binomial trial. Given the size of the movie n, the number of new actors follows a binomial distribution $\mathrm{BIN}(n, U)$, where U is a beta-distributed random variable, i.e. $U \sim \mathrm{beta}(p, q)$, and is independent of everything else. We are interested in a good model producing the number of new actors that look and feel like the real data. If we assume our model is true, then we obtain the maximal likelihood estimate (MLE) $\hat{p}$ and $\hat{q}$ for our proposed model. This is done by writing down the likelihood function and applying Newton-Raphson method. We obtain $\hat{p} \approx 0.2615$ and $\hat{q} \approx 2.097$.

We justify the model by comparing the outcomes of simulating the number of new actors with our proposed model to the real dataset. We did a simulation experiment of 27647 movies with the obtained MLE parameter and compare it with the real data

from all the movies in year 1994, as in Figure 5. We immediately see that the patterns are almost identical and hence come to that our model is a reasonable one.

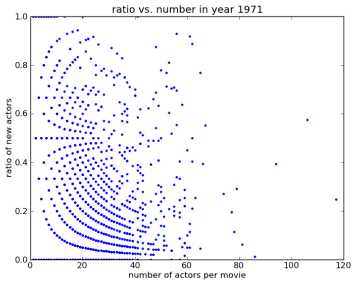
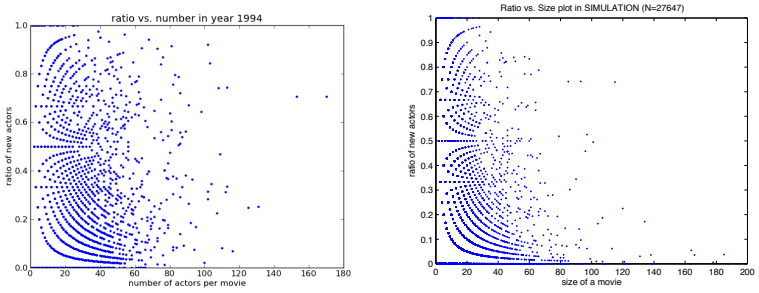


Figure 4: Ratio of new actors plot of movies in 1971



(a) Real Data in 1994 (b) Simulation of 27647 movies

Figure 5: Histogram Comparison

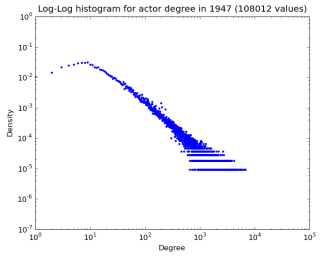
3.3 Preferential Attachment function

In this section we establish the preferential attachment function f by studying the evolution of actor degrees and movie degrees. If we think of a conventional collaboration network and try to fit a preferential attachment model there, we would often go for the linear dependence on the degrees. Nevertheless we have two kinds of degrees for our model – movie degrees and actor degrees – where movie degree of an actor is the number of movies the actor took part in and the actor degree is the number of other actors with whom the actor collaborated. The convention calls us to have a linear preference over the actor degree. If an actor has actor degree l, then the probability of this actor being chosen for a new movie is proportional to $l + \delta$ with δ extra flexibility parameter. We argue that in the dataset of the IMDb, it is better to have the linear preference over the movie degree. We study the actor and movie degree evolution separately and explain the reasoning.

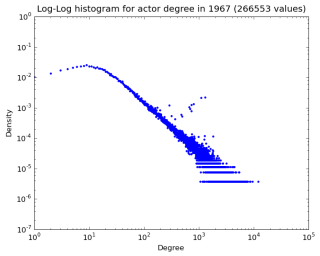
3.3.1 Actor degree evolution

We have a actor network given by the IMDb where the actors are vertices and the collaboration relationships are edges. The network evolves with time as new movies are produced and new actors are introduced into the network, resulting in the dynamics of the network. We study how the actor degrees evolves. We first present a plot of loglog-histogram of actor degrees in year 1947 in Figure 6. The four plots in Figure 6 are of accumulative nature, the actor network in year 1947 is a subset of that in year 1967 and so forth. Nonetheless the distribution of actor degrees is remarkably stable over 60-year time as we compare the four plots from different years, indeed this means the actor network is *scale-free*. It seems like the distribution of actor degrees follows a power-law. We perform a regression on the tail of loglog-histogram plot and, noting power-law is an asymptotic property, try to fit a straight line for actor degrees greater than 50, we obtain Figure 7 and an estimate of the power exponent $\tau \approx 2.0924$ for the actor network.

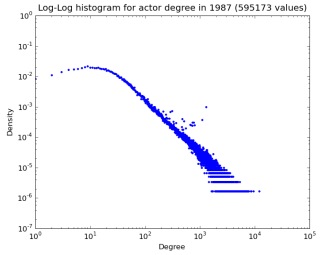
Note that if $p_k \propto k^{-\tau}$, $\log \sum_{j \geq k} f_j \sim -(\tau - 1) \log k$, thus $\log \big(1 - \hat{F}_N(x)\big)$-versus-$\log k$ plot gives a straight line with slope $-(\tau - 1)$. To further investigate whether the actor degree distribution follows a power-law, we do a $\log \big(1 - \hat{F}_N(x)\big)$-versus-$\log k$ plot in Figure 8. We see that for k not so large (before center of the plot horizontally) the curve looks like a straight line but goes downhill afterwards. This might suggest power-law is not exactly accurate for the actor degrees but some other modified power-laws suits it better. It is further discussed in Section 6.



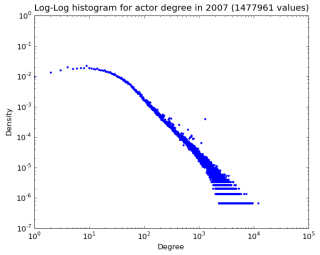
(a) 1947 with 108,012 actors



(b) 1967 with 266,553 actors



(c) 1987 with 595,173 actors



(d) 2007 with 1,477,961 actors

Figure 6: Actor degree evolution in 60 year from 1947 to 2007

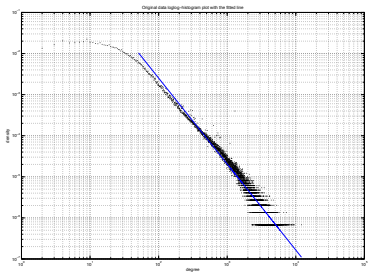


Figure 7: Fitting a straight line on the `loglog`-histogram starting from $k = 50$ on actor degrees

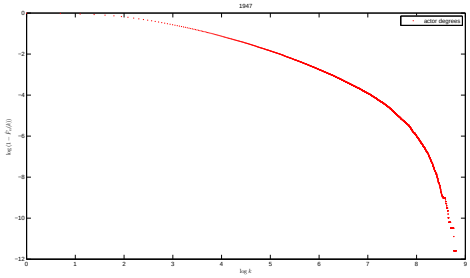
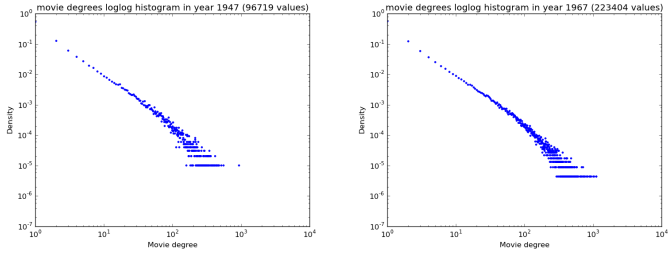


Figure 8: $\log\big(1 - \hat{F}_N(k)\big)$-vs.-$\log k$ plot of actor degrees in year 1947

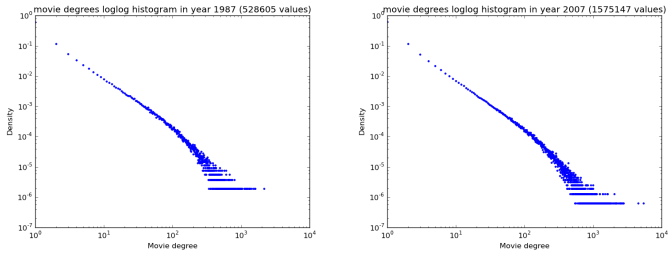
3.3.2 Movie Degree Evolution

We do the same for movie degrees as for actor degrees. Movie degrees and actor degrees are clearly closely related – an actor who participates in a lot of movies are, although not definitively, more likely to be linked with other actors and vice versa. Movie degrees over a 60-year time span is presented in Figure 9. Conclusions on movie degrees seem to converge with those of actor degrees – *scale free* and *power-law*. Same regression fitting, given in Figure 10, provides us with an estimate of the power-law exponent $\tau \approx 2.1654$.

Careful comparisons between Figure 6 and Figure 9 reveals that there are differences between movie degrees and actor degrees. The curve in the plots of Figure 9 is cleaner with less noise and seems more likely to be a straight line than the curve in Figure 6. This means the movie degrees are more stable and robust to random noise. This is understandable, since if there is a movie with a large size coming to the network, there are a lot of actors with high actor degree joining the network because all actors in the movie are linked with each other in the actor network.



(a) 1947 (b) 1967



(c) 1987 (d) 2007

Figure 9: Movie degree evolution in 60 year from 1947 to 2007

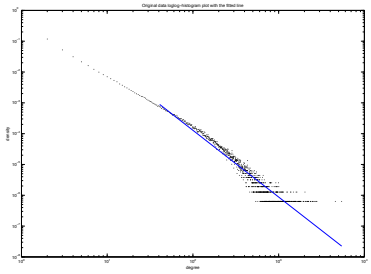


Figure 10: Fitting a straight line on the `loglog`-histogram starting from $k = 40$ on movie degrees

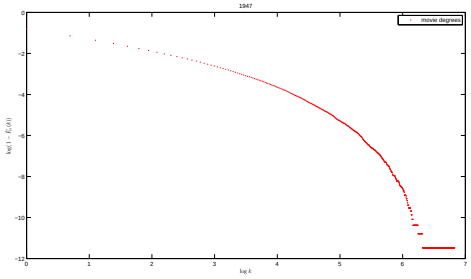


Figure 11: $\log\left(1-\hat{F}_N(k)\right)$ vs. $\log k$ plot of movie degrees in year 1947

3.4 Preferential Attachment Function

We see in the previous section that movie degrees are more stable and more robust against random noise, thus we choose to build the preferential attachment function upon the movie degree. Here we consider the preference to be a function of the movie degrees. Denote the weight function by $f(\cdot)\colon \mathbb{N} \to \mathbb{R}^+$, thus for (1) we make it more concrete

$$\mathbb{P}(a \leftrightarrow m_{t+1} | \mathfrak{G}_t) \propto f\left(D_t^M(a)\right) \quad \text{for any } a \in \mathcal{A}_t, \tag{2}$$

where $D_t^M(a)$ is movie degree of actor a at time t. In several recent papers Because of power-law exhibited in the distribution of movie degrees, we choose $f(k) = k + \delta$. In several recent papers [5, 12, 6, 2], different types of preferential attachment function were studies in three categories: (i) linear preferential attachment $f(k) \sim k$, (ii) sub-linear preferential attachment, $f(k) \sim k^\gamma$ when $0 \leq \gamma < 1$; (iii) super-linear preferential attachment $f(k) \sim k^\gamma$ when $\gamma > 1$. If it is the super-linear case, [6, 12] has shown that a "winner take all" phenomenon arises, where there is a single dominant vertex (often referred as "hub") linked to almost every other vertices. In [5], it is proved that a sub-linear preferential attachment results in an asymptotic degree distribution with a stretched exponential tails ([7]). We have pointed out that the distribution of movie degrees has thinner tails than a power-law, but ignoring the tail, the distribution of movie degrees gives a pretty good power law, which implies a linear preferential attachment model. Thus in sum, given what we have known about both the actor degree and movie degree distributions, the linear preferential attachment is reasonable here. Further discussion is in Section 6.

3.5 Model Fitting

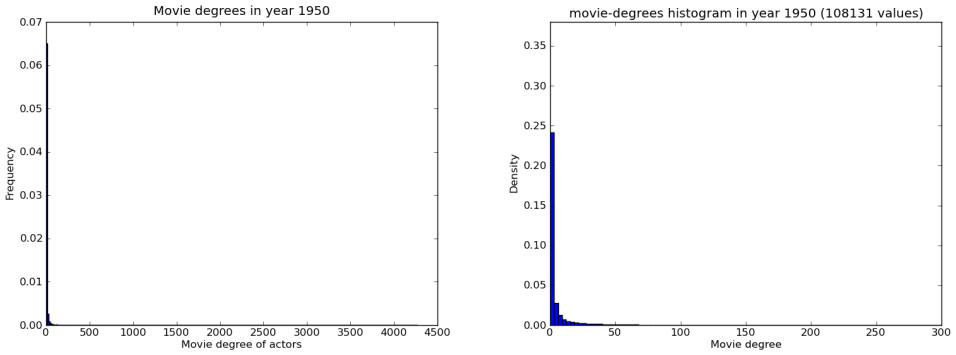
For ζ, we use the empirical distribution of movie sizes that we obtain from the real dataset. Given $\zeta = n$, we model the number of new actors as the randomized binomial experiment. With $U \sim \text{beta}(p, q)$ being independent, $\xi \sim \text{BIN}(n, U)$. Preferential attachment function is defined in (2). Hereafter the conceptual model filled with the above details is referred as the IMDb-PAM model.

4 Simulations

After fitting the model, it is natural to ask the question whether the model is good or not. We study how good the model is by doing simulations of the IMDb-PAM model and comparing the evolutions of the simulated network and network from real dataset.

It is worth noting that in our model t is not calendar time. It is easier if we somehow simulate the network in a yearly fashion because comparisons between the simulations and real dataset are easier over the years. We introduce a *"pseudo-year"* approach of generating the same number of movies as in the real dataset. For example there were 5624 movies in year 1915, then we simulate 5624 movies and mark them being in year 1915 and compare the respective networks at the end of year 1915. It is computationally costly to do the simulations as the number of actors blow out exponentially with the number of movies. Hence we only do the simulations until year 1951 with about $11,000$ movies in total and compare the respective networks in a 20-year interval (i.e. in 1910, 1930 and 1950). The author uses the same programming toolsets for the simulation and the analyses of simulations as for the real dataset.

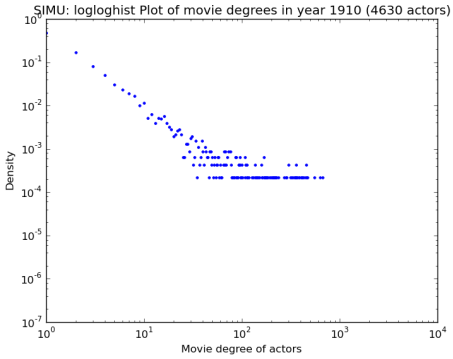
Movie degrees and actor degrees are two competing candidates upon which we could possibly build the preferential attachment, as stated in Section 3. In consequence we are mostly interested in the comparisons of these two quantities between the simulations and the real dataset. The comparisons are of interest as well for the reason that the two quantities are of (more or less) power-law distribution in the real dataset. We are concerned whether our proposed IMDb-PAM model recovers the power laws.



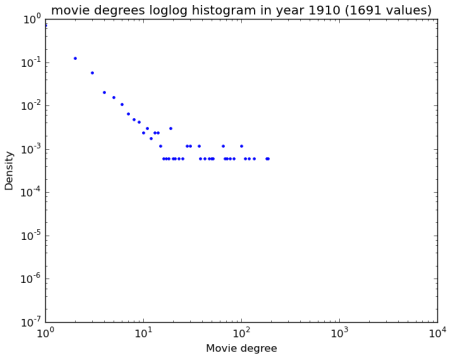
(a) Simulation (b) Real data

Figure 12: Histogram of actors' movie degrees by 1950

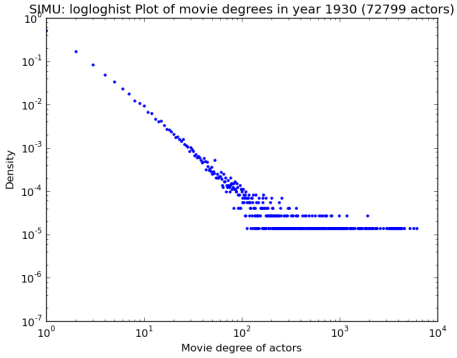
From Figure 12, we see that though of a similar shape, the histograms are different both in the horizontal axis (the degree) and the vertical axis (the frequency). However a better way of revealing the topological structure of degree sequences is loglog-plot of frequency versus degree. It discloses in particular whether the degrees follow a power-law distribution. We move to study the loglog histograms.



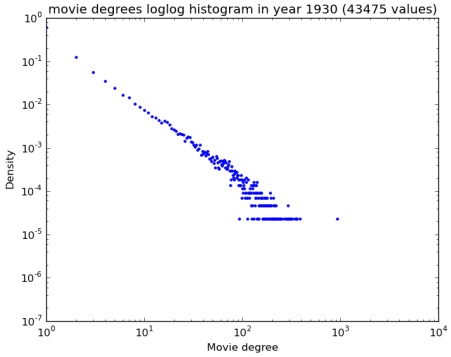
(a) Simulation until 1910



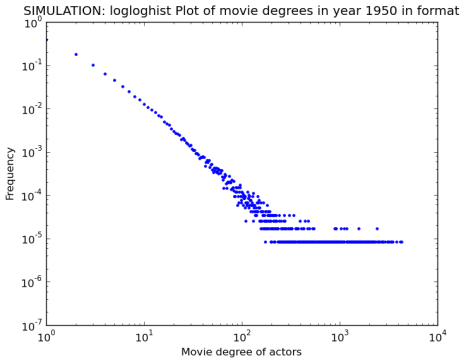
(b) Real data until 1910



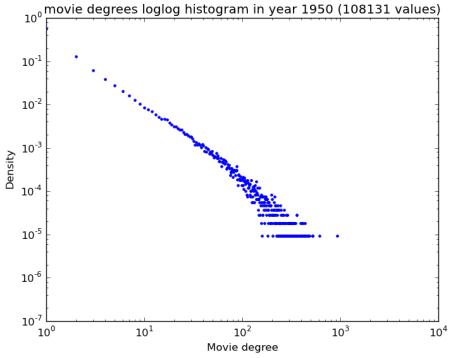
(c) Simulation until 1930



(d) Real data until 1930

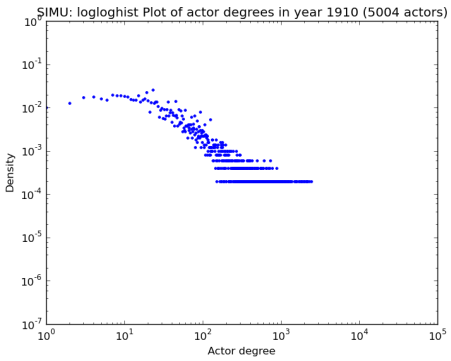


(e) Simulation until 1950

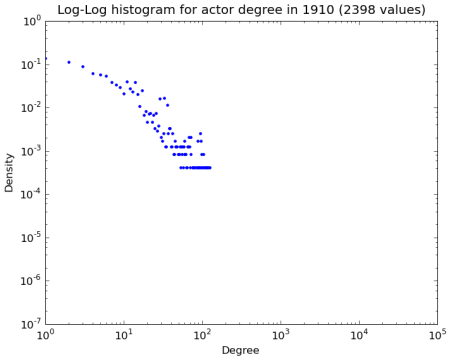


(f) Real data until 1950

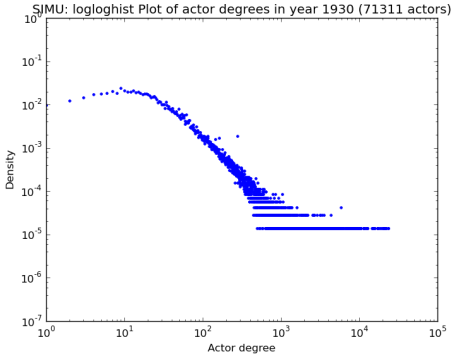
Figure 13: Movie degree comparisons between simulation and real dataset



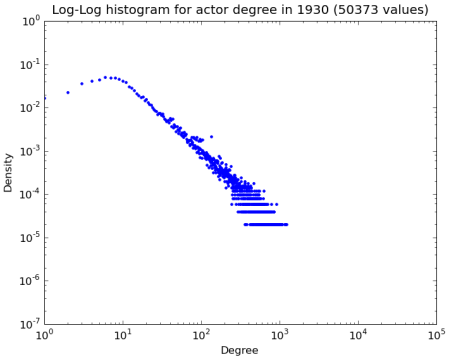
(a) Simulation until 1910



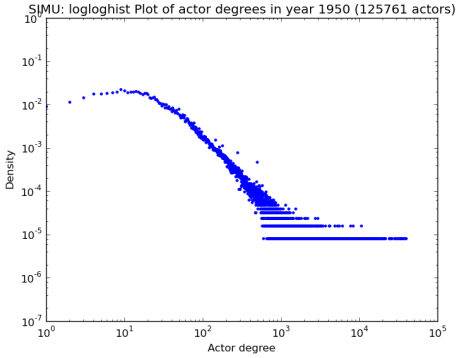
(b) Real data until 1910



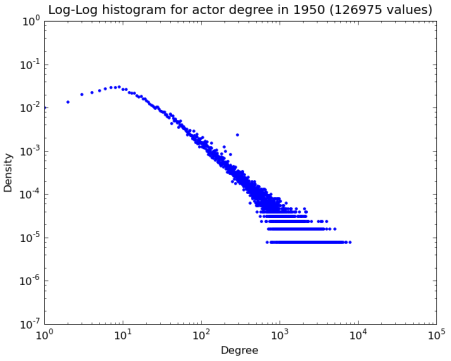
(c) Simulation until 1930



(d) Real data until 1930



(e) Simulation until 1950



(f) Real data until 1950

Figure 14: Actor degree comparisons between simulation and real dataset

Figure 13 and Figure 14 are comparisons of movie degree evolution and movie degree evolution over 50 years of time span. We read from these plots that for both actor degree and movie degree, the fit of the simulation to the real dataset is quite good but with a certain degree of imperfectness. The simulation gives a power-law distribution with similar power-exponent but there are more variations for large degrees in simulation than in real data. The tail in the simulation is heavier, indicating that there are more actors with high degrees. The maximal degrees, both of movie degree and actor degree, are one order of magnitude higher in the simulation than those in the real dataset. Nonetheless the simulation outcome is not entirely same for the actor degree and movie degree: the fit of movie degree seems better than that of actor degree the difference between slopes of curve in the tail in the simulation and real dataset is bigger for actor degrees.

The above observations, on one hand, suggests that our model is basically a reasonable one. Indeed the simulation gives quite accurate, if not perfect, capturing of most actors' degrees. On the other hand, the simulation exposes the systematic error of our IMDb-PAM model because the simulation gives distinctive characteristics for the tails of the degree sequence. The IMDb-PAM model is putting more preference on the actors with higher degrees than the real dataset. It is perhaps reasonable that one should put some restriction on choosing the actors with high degrees. After all one actor cannot play in too many movies for the physical constrains. Possible remedies are mentioned in Section 6.

5 Theoretical Study

We present a theorem without giving the proof. For those interested, the proof is given separately as attachment to this paper. The theorem ensures that our movie-actor model gives us the desired asymptotic power-law movie degree distribution. We introduce the notation $\mu_\psi = \mathbb{E}\psi$, $\mu_\zeta = \mathbb{E}\zeta$ and $\mu_\xi = \mathbb{E}\xi = \mu_\zeta - \mu_\psi$. To simplify the notation we also need $\theta = \mu_\zeta + \mu_\xi\delta$ and $\theta^* = \theta/\mu_\psi$.

Theorem 1. *Assuming that the followings hold:*

1. *There exist a constant $a_0 > 0$ and c_N such that*

$$\mathbb{P}(\zeta > N) \leq c_N N^{-(3+a_0)}. \tag{3}$$

   In particular this implies the distribution of the movie size ζ has a finite second moment.[1]

2. $\delta > -1$.

3. $\theta > 1$.

[1] In fact this implies the distribution of movie size ζ has a finite third moment, but a finite second moment is sufficiently strong to proceed the proof.

Then there exists a constant γ such that

$$\lim_{t\to\infty} \mathbb{P}\left(\max_{k\geq 1} |p_k(t) - p_k| \geq t^{-\gamma}\right) = 0$$

where $(p_k)_{k\geq 1}$ is defined as the solution to the recursive equation $p_k = \frac{k-1+\delta}{\theta^*} p_{k-1} - \frac{k+\delta}{\theta^*} p_k + \mathbf{1}_{\{k=1\}}$ for $k \geq 0$.

The recursive equation is solved by $p_k = \frac{\Gamma\left(1+\delta+\theta^*\right)}{\Gamma(1+\delta)} \frac{\Gamma(k+\delta)}{\Gamma\left(k+\delta+1+\theta^*\right)} \theta^*$. In particular $(p_k)_{k\geq 1}$ follows a power-law as $p_k \approx c(\theta^*, \delta) k^{-(1+\theta^*)}$ when k is sufficiently large.

6 Conclusion and Future work

We have proposed a model and fit our model to the specific case of the IMDb dataset. The simulations have shown that our model is a good fit hence our model provides insight on the generative mechanism how the present actor networks come into being. However the model is not of perfectness and is subject to substantial improvement.

We could consider a saturation of power-law – power-law with exponential cut-off. A standard power-law is of the form that $p_k = Ck^{-\tau}$ for some approperiate positive constant C and τ, if we consider the form $p_k = Ck^{-\tau} \exp^{-k/A}$ for some approperiate constant C, τ and A. For the latter form $p_k \approx Ck^{-\tau}$ when $k \ll A$ but the exponential term is dominating when k is sufficiently large, hence $p_k \propto exp(-k/A)$. If we do a $\log(1 - \hat{F}_N(k))$-versus-$\log k$ plot for the data drawn from a distribution of a power-law with exponential cut-off, we see curve with more-or-less straight line until k is moderately large and then a concave curvature with slope decreasing rapidly, which is exactly what we see in Figure 8 and Figure 11. Invesigating how to fit such an model into the dataset of the IMDb might be interesting and so is looking into mechanisms that are responsible for the cut-off.

A different preferential attachment model can be pursued as well. Instead giving linear preference with respect to degrees, we consider a truncated linear preferential attachment function in (2)

$$f(k) = \begin{cases} k+\delta & \text{for } k \leq A, \\ f_1(k) & \text{for } k \geq A, \end{cases} \tag{4}$$

where $f_1(k) : \mathbb{N} \to \mathbb{R}^+$ is a function that grows much slower than a linear function, or even a constant function. The truncated preferential attachment, on one hand, accommodates the assumption that preference should be given to people with higher degrees; on the other hand restricts the preference by the truncation where the restriction is due to physical constrains and so on.

If we assume little knowledge on the preference attachment function, it is interesting to devise a statistical methodology to estimate f upon imposing some weak a priori assumption on f, although difficult as it is nonparametric problem. The author is pursuing the objective as an independent project within the framework of Bayesian nonparametric statistics.

Acknowledgement

The paper is based on the corresponding author's master thesis in Eindhoven University of Technology. The project was supervised by Remco van der Hofstad and Rui Castro. The original programming toolset of analysing the dataset was written by Vincent Kusters, who was then a student in Eindhoven University of Technology. After the master project, the corresponding author went on doing a PhD with Aad van der Vaart in Leiden University.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Sagy Bar, Mira Gonen, and Avishai Wool. An incremental super-linear preferential internet topology model. In *Passive and Active Network Measurement*, pages 53–62. Springer, 2004.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] Maria Deijfen, Henri van den Esker, Remco van der Hofstad, and Gerard Hooghiemstra. A preferential attachment model with random initial degrees. *Arkiv för matematik*, 47(1):41–72, 2009.

[5] Steffen Dereich and Peter Mörters. Random networks with sublinear preferential attachment: Degree evolutions. *Electron. J. Probab*, 14(43):1222–1267, 2009.

[6] Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[7] Jean Laherrere and Didier Sornette. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[8] Mark EJ Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131, 2001.

[9] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[10] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[11] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[12] Roberto Oliveira and Joel Spencer. Connectivity transitions in networks with super-linear preferential attachment. *Internet Mathematics*, 2(2):121–163, 2005.

[13] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.