International Journal of Bifurcation and Chaos, Vol. 22, No. 7 (2012) 1250180 (12 pages)

DOI: 10.1142/S0218127412501805

MODELING THE EVOLUTION OF ITEM RATING NETWORKS USING TIME-DOMAIN PREFERENTIAL ATTACHMENT

EDMUNDO F. LAVIA
Physics Department, School of Sciences, University of Buenos Aires, Av. Cantilo s/n, Pab.1, Ciudad Universitaria, Buenos Aires 1428, Argentina

ARIEL CHERNOMORETZ
Physics Department, School of Sciences, University of Buenos Aires, Av. Cantilo s/n, Pab.1, Ciudad Universitaria, and Fundación Instituto Leloir, Av. Patricias Argentinas 440, and CONICET, Av. Rivadavia 1917, Buenos Aires 1428, Argentina

JAVIER M. BULDÚ
Complex Systems Group, URJC, 28933 Móstoles, Spain
Laboratory of Biological Networks, Centre for Biomedical Technology (UPM), 28922 Pozuelo de Alarcón, Madrid, Spain

MASSIMILIANO ZANIN
Centre for Biomedical Technology, Polytechnic University of Madrid, Pozuelo de Alarcón, 28223 Madrid, Spain
Faculdade de Ciências e Tecnologia, Departamento de Engenharia Electrotécnica, Universidade Nova de Lisboa, Portugal
Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006, Madrid, Spain

PABLO BALENZUELA
Physics Department, School of Sciences, University of Buenos Aires, Av. Cantilo s/n, Pab.1, Ciudad Universitaria, and CONICET, Av. Rivadavia 1917, Buenos Aires 1428, Argentina
balen@df.uba.ar

Received December 15, 2011; Revised March 26, 2012

The understanding of the structure and dynamics of the intricate network of connections among people that consumes products through Internet appears as an extremely useful asset in order to study emergent properties related to social behavior. This knowledge could be useful, for example, to improve the performance of personal recommendation algorithms. In this contribution, we analyzed five-year records of movie-rating transactions provided by Netflix, a movie rental platform where users rate movies from an online catalog. This dataset can be studied as a bipartite user-item network whose structure evolves in time. Even though several topological

properties from subsets of this bipartite network have been reported with a model that combines random and preferential attachment mechanisms [Beguerisse Díaz *et al.*, 2010], there are still many aspects worth to be explored, as they are connected to relevant phenomena underlying the evolution of the network. In this work, we test the hypothesis that bursty human behavior is essential in order to describe how a bipartite user-item network evolves in time. To that end, we propose a novel model that combines, for user nodes, a network growth prescription based on a preferential attachment mechanism acting not only in the *topological domain* (i.e. based on node degrees) but also in *time domain*. In the case of items, the model mixes degree preferential attachment and random selection. With these ingredients, the model is not only able to reproduce the asymptotic degree distribution, but also shows an excellent agreement with the Netflix data in several time-dependent topological properties.

Keywords: Complex networks; bipartite networks; Netflix; bursting; recommendation systems.

1. Introduction

During the past years, we have witnessed a wide range of contributions on the applications of Complex Networks Theory to real data [Newman, 2003; Boccaletti *et al.*, 2006; Costa *et al.*, 2011]. The main reason behind this explosion is the large number of datasets accessible to any user through Internet. Nevertheless, an excess of information can, sometimes, be a disadvantage, since a user may have problems to find specific information or even get lost in the pool of data. Personal recommendation algorithms deal with this drawback of large datasets, and have been specially fruitful in the context of music [Herlocker *et al.*, 2004; Cano *et al.*, 2006; Zanin *et al.*, 2009; Celma, 2010] or movie recommendation [Zhang *et al.*, 2007; Zhou *et al.*, 2007; Beguerisse Díaz *et al.*, 2010]. Within this framework, collaborative filtering methods [Sarwar *et al.*, 2001; Herlocker *et al.*, 2004] have shown very high performance as measured by high scores in their recommendation results. This kind of algorithms rely on the data previously collected from users' behavior, namely the number, type and rating of the items they have consumed. In the last decade, a lot of effort has been made in order to improve collaborative filtering algorithms, trying to increase their score in the prediction of what users' next choice will be [Bobadilla *et al.*, 2009; Zanin *et al.*, 2009]. Nevertheless, less attention has been paid to the data that these recommendation algorithms are using as a ground for their automatic predictions. In the current work, we are concerned about the creation and evolution of rating networks, which are usually taken as the input of collaborative filtering algorithms. Rating networks are bipartite networks [Holme *et al.*, 2003] whose fundamental nodes are split into two kinds, users and items, and links are created when a user gives a rate to a certain item that he/she has consumed. In this way, we obtain (complex) rating networks that are continuously evolving in time, increasing its number of users, items and links. Due to its relevance and availability, we have analyzed the rating dataset given by Netflix, an online movie rental platform [Netflix, 2011].

Statistical features of Netflix data were already the subject of several studies. Some of them have focused on collaborative filtering procedures and/or recommendation algorithms [Bennett & Lanning, 2007; Zanin *et al.*, 2009], others [Beguerisse Díaz *et al.*, 2010] in describing relevant topological properties of the subsets of the Netflix database. In the present contribution our aim is twofold: on one hand, we want to understand the underlying rules that drive the evolution of this rating network and, on the other hand, we want to design a model able to reproduce the main features of users, items and links. We will see that the analysis of one-year long top-rated movies shows a power-law distribution in the degree of items (movies) and an absence of a power law behavior in the degree distribution of users, as well as a non-Poissonian distribution in human activity time domain, which is characterized by bursts of intense activity (high number of ratings) followed by periods of inactivity. These long-tailed distribution was already reported for the Netflix network [Beguerisse Díaz *et al.*, 2010], as well as for other kinds of human activities [Barabási, 2005; Oliveira & Barabási, 2005; Vazquez *et al.*, 2006, 2007; Zhou *et al.*, 2008].

Interestingly, we found that a model for the creation of new links that is only based on preferential attachment and random selection is not capable of reproducing the complete abovementioned observations. In particular, it is not able to reproduce the bursty behavior observed in user node's dynamics. In this contribution, we will show that a combination of preferential attachment in degree and time-domain is indeed needed to better describe the evolving rating networks. Moreover, the observed balance between these driving forces could also be taken into account in order to design efficient recommendation algorithms based on this kind of datasets.

The manuscript is organized as follows: in Sec. 2, we define the main properties of the Rating Network (RN) we are studying; in Sec. 3, we show the results obtained when analyzing the network structure and its temporal evolution; Sec. 4 is devoted to the design of an evolutionary model that reproduces the fundamental properties of the Netflix Rating Network; finally, in Sec. 5, we summarize the results obtained and discuss the implications that the network evolution has on the design of recommendation algorithms.

2. The Netflix Rating Network

Netflix is an online movie rental platform upon which a social network of user-assigned video ratings was established. Six years of online ratings (i.e. transactions) were made publicly available as a part of the Netflix competition [Bennett & Lanning, 2007] in the year 2007. The whole dataset includes a collection of 480 189 users, 17 770 movies, and 100 480 507 ratings, spanning about six years of activity, from October 1998 to December 2005.

The Netflix rating dataset can be naturally cast into a bipartite network representation in which users and movies are continuously entering the system. New links are established between both types of nodes every time a rating transaction is registered. A schematic picture of the network temporal dynamics can be seen in Fig. 1.

The first user's rating determines his/her entry time point to the market. Analogously, the first rating a given film receives defines its entrance in the network. These two quantities can be used to trace the overall dynamical evolution of the Netflix expanding market. From the observed dynamics, phenomenological growing laws, $M(t)$ and $U(t)$, can be inferred for movies and users respectively (see

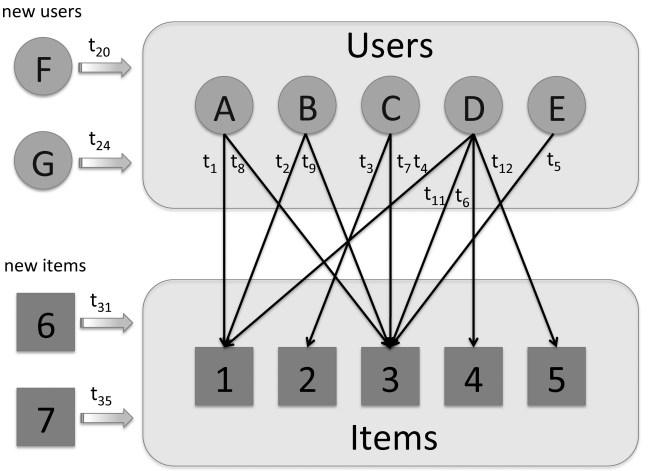


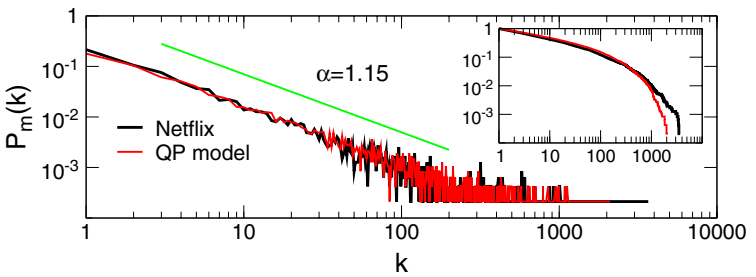
Fig. 1. Construction and evolution of a user-item rating network. Users, items and ratings appear at discrete times t_i. A new link is created when a user rates a certain item (movies, in the case of Netflix). Users and items are continuously added to the system. In this qualitative example, Item 3 would be a network hub.

Fig. 7 in Appendix A). In the following sections we have made use of this phenomenological growing curves in order to model the network dynamics.

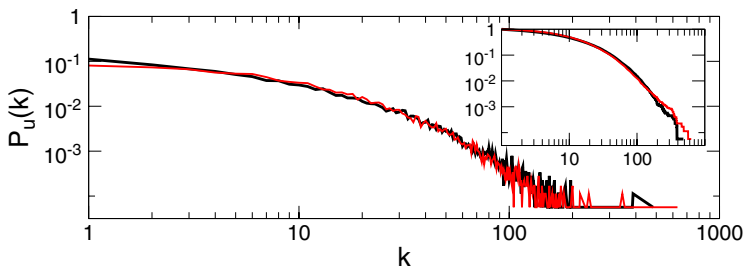
For the sake of computational modeling efforts, we decided to consider only a representative subset of the complete database. On one hand, we focus on ratings that were worth the highest score (five stars) in order to assure that the user's feeling about the movie is fully positive. On the other, we only keep transactions that have occurred between January 6, 2001 and January 6, 2002. We verify that several topological and dynamical features remain unaltered for other date choices, as long as a whole year of sampling period was considered (see Fig. 8 in Appendix A). This is somehow to be expected, as a calendar one-year period can be considered a natural time scale to describe human-related activity patterns. Keeping track of five-stars transactions over one year resulted in a network of $U = 17\,729$ users, $M = 4734$ movies and $T = 300\,351$ ratings (links).

3. Topological Properties of Netflix Bipartite Network

The first step in order to unveil a bipartite network connectivity pattern can be done by analyzing the degree distribution of user and movie nodes. They are depicted in Fig. 2, in panel (a) for movies and panel (b) for users. The movie's degree distribution displays a dominant power-law behavior



(a)



(b)

Fig. 2. Degree distributions ($P(k)$) for (a) movies and (b) users. Netflix data from 2001 (in black) as well as results of QP model for $Q = 0.56$ and $P = 0.99$ (in red) are shown. Meanwhile, the one corresponding to movies displays a predominant power law behavior with exponent $\alpha \sim 1.15$ in the range ($k = 1$–300), the distribution corresponding to users shows a strong exponential character. The insets show the cumulative degree distributions.

(truncated by finite size effects) with an exponent of $\alpha \sim 1.15$, although this is not the case for the user's degree distribution. The simplest assumption in order to model this behavior is that a preferential attachment mechanism underlies the selection procedure of movies, meanwhile a mixing with random selection should be involved in the users's behavior. This approximation was already used in [Beguerisse Díaz *et al.*, 2010] and it will be tested in Sec. 3.1.

3.1. *A basic approximation: Preferential attachment and random selection*

We start with a simple model (the QP-Model), similar to the one presented in [Beguerisse Díaz *et al.*, 2010], which combines preferential attachment and a random node selection prescription onto an evolving user-item network. The dynamics of the model is defined by the evolution of the nodes and the rules by which a user selects a given item. The evolution in time of movies and users, $M(t)$ and $U(t)$, are taken as empirical growth laws, since they were fitted from the data (see Appendix A for details). Each simulation step, t_s, corresponds to a rating transaction that links a user to a movie. If a new user (movie) node has to be incorporated to the market at a given simulation step, it would be the one selected to participate in the transaction. Otherwise an already existing user (movie) node would be selected following the $P(u_i, t_s)$ ($P(m_i, t_s)$) probability distribution function:

$$P(u_i, t_s) = Q\frac{k_i(t_s)}{\sum_l^{U(t_s)} k_l(t_s)} + (1-Q)\frac{1}{U(t_s)} \quad (1)$$

$$P(m_j, t_s) = P\frac{k_j(t_s)}{\sum_l^{M(t_s)} k_l(t_s)} + (1-P)\frac{1}{M(t_s)}, \quad (2)$$

where $U(t_s), M(t_s)$ are the number of users and movies that are already in the market after t_s transactions and $k_{i/j}$ is the node degree. $Q \in [0, 1]$ and $P \in [0, 1]$ are model parameters that control the relative strength between the preferential attachment and the random assignment character of the probabilistic assignment rule for users and movies, respectively. The entire simulation included $M = 4734$ movies, $U = 17\,729$ users and $T = 300\,351$ ratings which corresponds to a one-year period (2001).

It is important to point out that even though human days (labeled as t in this work) were the time units used in the original dataset, we considered the number of transactions (t_s) as the model time scale. The correspondence between days and transactions was fitted from the original dataset following the same approach used in [Beguerisse Díaz *et al.*, 2010] (see Appendix A for further details).

As can be seen in Fig. 2 (in red), the model could adequately fit the empirical data for the degree distribution of both type of nodes ($Q = 0.56$, $P = 0.99$, $\chi^2_{\mathrm{NORM}} = 0.14$). Details about the fitting procedure can be seen in Appendix A. Looking at the best fit parameter values it can be realized that rather large deviation from a pure preferential attachment behavior is obtained for user nodes, given the empirical node degree distribution which displays a strong exponential character.

3.2. *Network dynamical features*

Given that the network is continually growing with the influx of new nodes (users and movies), it is sensible to investigate some dynamical features of the system in order to better understand its temporal organization. The trivial tendency of elder nodes to have participated in more transactions than newcomers, just because they were around longer in the market, was analyzed in Figs. 3(a) and 3(b). In these panels we show the average node degree $\langle k \rangle$ as a function of the insertion date, for users and movies, respectively. It is remarkable that, whereas older films tend to show higher average degree values than recently incorporated ones, no such strong correlation could be established for users. This asymmetry could not be recapitulated by the QP-model (in red).

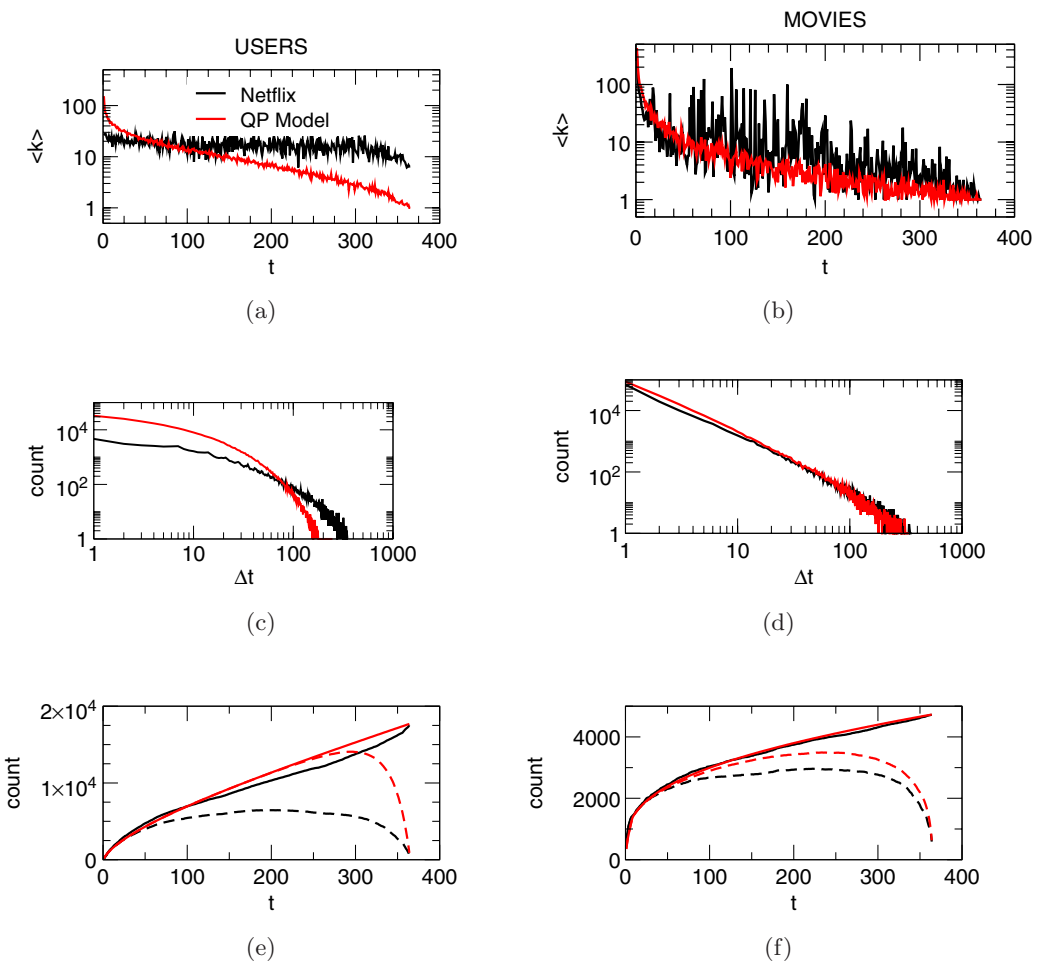


Fig. 3. Topological dynamical analysis of the Netflix data (in black) and QP-Model ($Q = 0.56$ and $P = 0.99$, red) for (left) users and (right) movies. (a) and (b) Mean degree $\langle k \rangle$ as a function of the market insertion date, (c) and (d) distribution of times between transactions, $P(\Delta t)$, (e) and (f) the life cycle of users and movies. Active (total) nodes are shown with dashed (full) lines. Left panels show how the QP model fails to describe the dynamical features of the network.

The rate of individual transactions is another interesting dynamical feature of our system. Figures 3(c) and 3(d) show the distribution of times between transactions, $P(\Delta t)$, for users and movies. Aside from the expected seven-day pattern of activity (small ripples in the figure, also in panel (c) of Fig. 8) which corresponds to typical inter-event period in time organization of many human activities, we found a heavily-tailed distribution for both, users and movies distributions, with a stronger power-law dominance in the latter case. We can observe that the QP-model (red lines) could fit accurately enough the data corresponding to movies. However, also in this case, it fails to reproduce the observed behavior for users, displaying a larger than observed fraction of small, and midsize inter-event intervals, and underrepresented largely delayed patterns of activities.

In the same figure, in panels (e) and (f), the permanence of users and movies in the network is analyzed. We define the duration of the spanning life cycle of a node over a finite temporal window, as the number of days that mediate between its appearance on the network (first rating recorded) and their last rating within the analyzed period. We consider that a node is active if it has not reached its corresponding final degree. Again, and in concordance with the above observations, we noticed that the QP-model fails to mimic the observed behavior for users. The overrepresentation of small inter-event times, observed in panel (c) for user transactions results in the consumers with similar degrees remaining active for longer times in the model than in the real network.

4. A New Model: Degree and Time Preferential Attachment

At this point, it becomes evident that even the node degree distribution could be well adjusted by the QP-model, the preferential-attachment and random selection mechanisms did not convey the model enough flexibility to adequately fit the reported dynamical behavior of the system. Even if several movie-related temporal observables could be nicely adjusted by the model, this was not generally the case of user-related dynamical behavior. This kind of qualitative asymmetry could not be corrected by different values of parameters Q and P. On the contrary, it reflects the intrinsic different nature between both types of nodes, and the complexity of human temporal task organization. In order to look for differences in temporal patterns between users and movies we plot, in Fig. 4, the time interval between consecutive transactions as a function of the transaction number for randomly selected users and movies of three arbitrarily categories: highly connected nodes ($k \sim 360$ for users $k \sim 2100$ for movies), regular ones ($k \sim 155$ for users $k \sim 615$ for movies) and low connected nodes ($k \sim 50$ for users $k \sim 110$ for movies).

Figure 4 contrasts the dynamical activity of movies and users. It can be noticed that the latter ones display bursts of activity separated from long periods of inactivity, consistently with reported patterns of inter-event distribution associated with human dynamics [Barabási, 2005; Oliveira & Barabási, 2005; Vazquez *et al.*, 2006].

In order to take into account these observations, we develop a new model of network evolution (the RP-model), which combines for user dynamics a preferential attachment in the degree with a preferential attachment in the time domain. We use the same empirical growing laws than in the QP-model (i.e. $M(t)$ and $U(t)$ as shown in Fig. 7) and the same number of movies, users and transactions ($M = 4734$, $U = 17\,729$ and $T = 300\,351$, respectively) but we change the probability of selecting an existing user at a given time. In this model, the probabilities of selecting a user or a movie are read as,

$$P(u_i, t_s) = R\frac{k_i(t_s)}{\sum_{l}^{U(t_s)} k_l(t_s)} + (1-R)\frac{\frac{1}{t_s - t^L_{s,i}}}{\sum_{l}^{U(t_s)} \frac{1}{t_s - t^L_{s,l}}} \tag{3}$$

$$P(m_j, t_s) = P\frac{k_j(t_s)}{\sum_{l}^{M(t_s)} k_l(t_s)} + (1-P)\frac{1}{M(t_s)}, \tag{4}$$

where $U(t_s), M(t_s)$ are the number of users and movies that are already in the market after t_s transactions, $k_{i/j}$ is the node degree and $t^L_{s,i}$ is the time step where the ith user has made his last rate. $R \in [0,1]$ is the model parameter that controls the relative strength between the preferential attachment in degree and preferential attachment in time domain. $P \in [0,1]$ plays the same role as in the QP-model.

This model takes into account the asymmetry observed in the temporal behavior of humans and

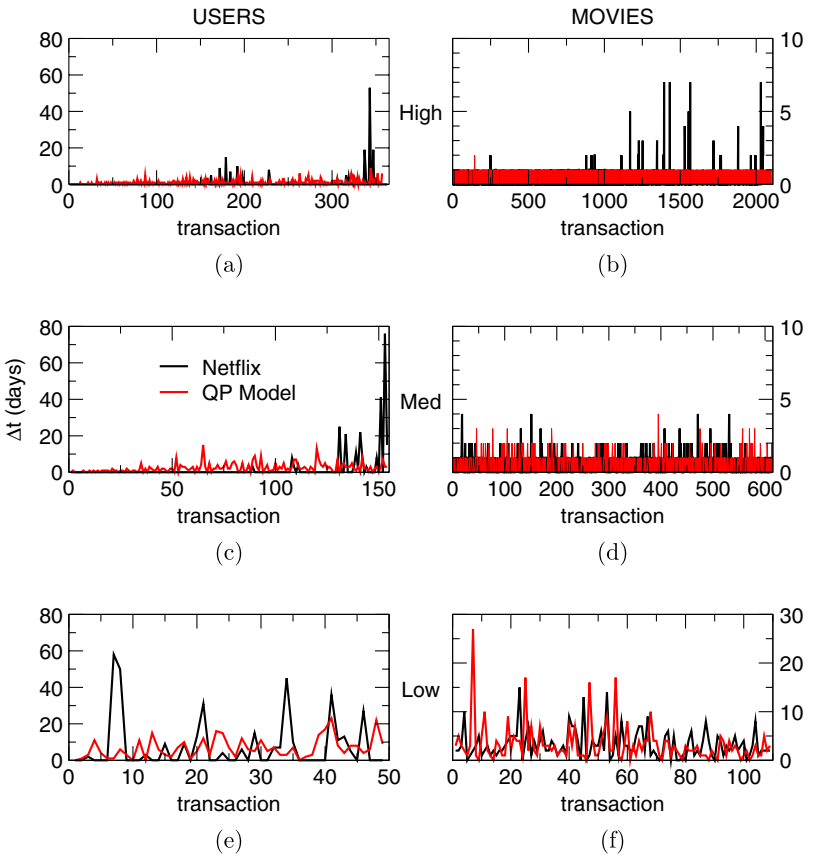


Fig. 4. Dynamical behavior of consumers and movies. The time interval between consecutive transactions as a function of the transaction number for (left) users and (right) movies are plotted. (a) and (b) High, (c) and (d) medium, (e) and (f) low connected nodes are sketched. Both columns highlight the differences in their dynamical behavior: Whereas the human activity is organized in burst of activity separated with long periods of inactivity for different kind of consumers, the movie's behavior depends on their degree of popularity. In particular, the most ranked movies are consumed almost every day. This plot also clearly shows how the QP-model fails to reproduce the behavior of users, but accurately describes how the movies behave.

movies by breaking the symmetry between nodes and movies selection probability rules. The second term of Eq. (3), that is, the one proportional to $1/(t_s - t^L_{s,i})$ ensures a succession of consecutive ratings for users who have recently qualified while the first term allows a user who has not rated for some time to requalify and enter again in a regime of bursts. In addition, the selection introduced by the first term of Eq. (3) is through the usual degree-based preferential attachment, and it promotes bringing back users of high degree.

Following the same procedure used with the previous model, we adjust the parameters R and P in order to get the best fit to the users and movies degree distributions in the year 2001 Netflix database. We obtained $R = 0.11$ and $P = 1.0$ with $\chi^2_{\rm NORM} = 0.25$.

In Fig. 5, we show the performance of the new RP-model (for $R = 0.11$ and $P = 1.0$), matching several topological/dynamical features of the Netflix network that were already examined in Figs. 2 and 3 for the simpler QP-model, i.e. the degree distributions [panels (a) and (b)], the mean degree as a function of the market insertion date [panels (c) and (d)], distribution of times between transactions, $P(\Delta t)$ [panels (e) and (f)], and the life cycle of users and movies [panels (g) and (h)]. We can appreciate how the incorporation of preferential attachment in time domain, instead of a random selection, is enough to qualitatively reproduce the main dynamical aspects of Netflix bipartite network.

For users, their dynamics in the RP-model is much closer to those observed in the data when we use the probability of the form Eqs. (3) and (4). Equiprobable random assignment, as in the QP-model, overemphasizes the importance of the permanence of the users so that, on average, only the

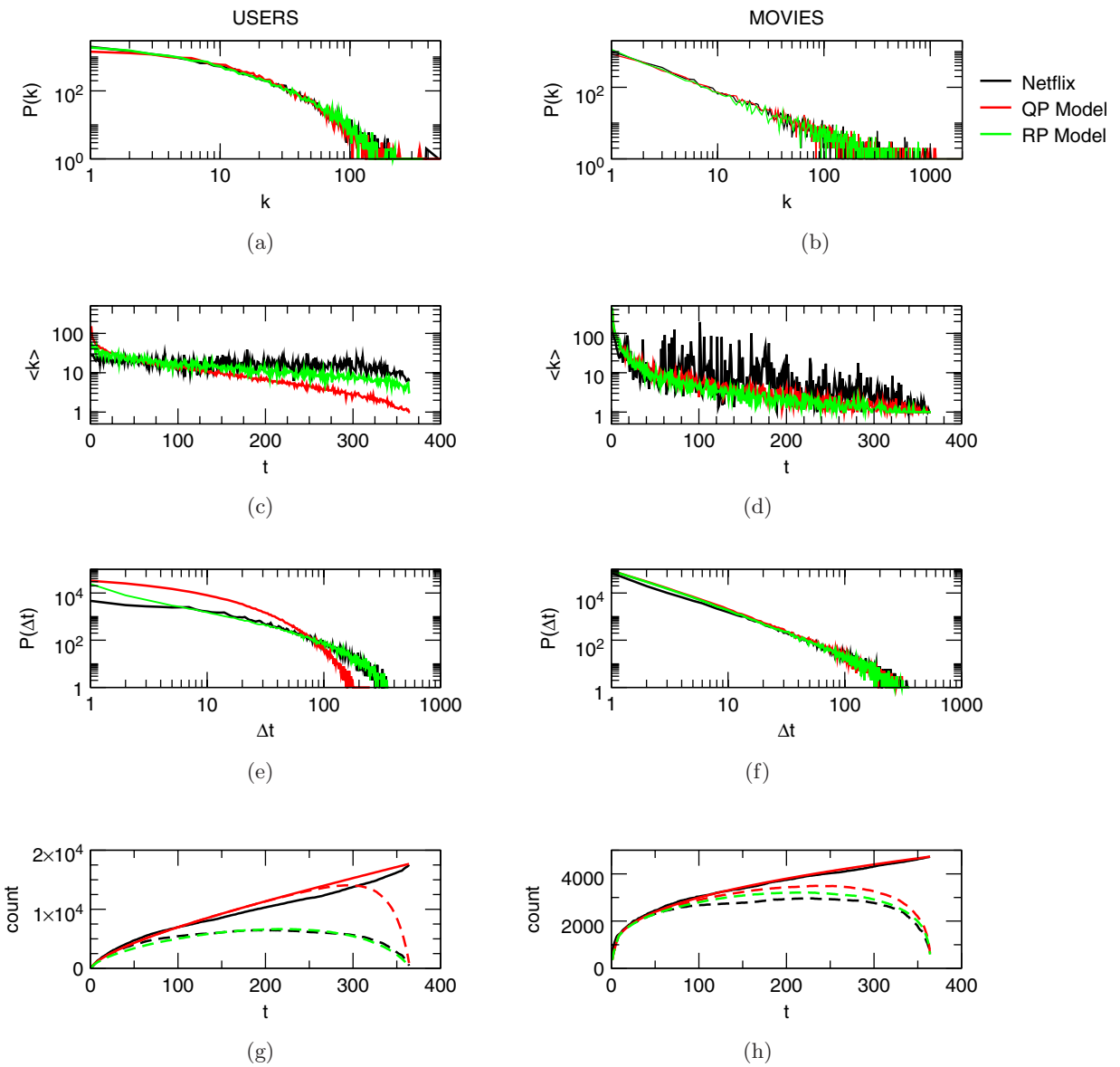


Fig. 5. Topological/dynamical analysis of the Netflix data (black), RP-Model ($R = 0.11$ and $P = 1$, green) and QP-Model ($Q = 0.56$ and $P = 0.99$, red) for (left) users and (right) movies. (a) and (b) Degree distributions, (c) and (d) mean degree $\langle k \rangle$ as a function of the market insertion date, (e) and (f) distribution of times between transactions, $P(\Delta t)$, (g) and (h) the life cycle of users and movies. Active (total) nodes are shown by dashed (full) lines. Left panels confirm that RP model describes accurately the dynamical features of the Netflix bipartite network.

initial nodes can be high degree nodes and also fails to capture the inter-event dynamics. The exponential character of the distribution, when using randomness, would indicate that the appearance of users follows a pattern more in accordance with a Poissonian process [Vazquez *et al.*, 2006] in which a time scale for time intervals between consecutive transactions can be defined.

In order to further characterize the temporal activity patterns generated by the RP-model and compare with those observed in Netflix data, we define an observable which quantitatively characterizes the user-nodes bursting behavior. For each user node in the model, we considered the sequence of inter-event intervals $\{\Delta t_j\}$ ($j \in [1, k-1]$ (as plotted in Fig. 4), where k is the node degree). We sorted this set in a decreasing order, and calculated the time, t_{90}, at which the cumulative inactivity period reaches 90% of the life time of each node. A small t_{90} value corresponds to a situation where the temporal transaction history of a given user is dominated by large inactivity periods.

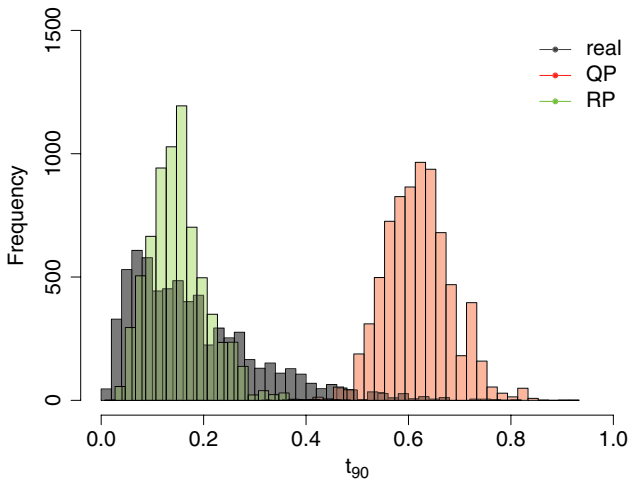


Fig. 6. Characterization of bursting behavior through inactivity periods of time. t_{90} distribution (normalized by the time at which the cumulative inactivity period reaches 90% of the life time of each user) is obtained for all user nodes of the real network (black), QP-model (red), and RP-model (green). Lower values of t_{90} indicate longer period of inactivity.

On the contrary, a user behavior dominated by small intervals will display a large value of t_{90}.

In Fig. 6, we report the t_{90} distribution obtained for all user nodes of the real network (black), QP-model (red), and RP-model (green). We can see that for real users, the t_{90} distribution is dominated by small values. This is compatible with the idea that the respective temporal dynamics shows a combination of a few large inactivity periods mixed with burst activity patterns. The curves associated to the RP-model present a rather similar character to the real ones (*albeit* they do show less variability than the real data). Finally, the t_{90} distribution of the QP-model indicates a behavior dominated by small time intervals. This is a remarkable result, as the model parameters were obtained just by fitting a static observable, i.e. the node degree distribution.

5. Conclusions

During the last years, rating networks have been a useful source of information for the development of personal recommendation algorithms. Nevertheless, the structure and evolution of this kind of networks has to be taken into account for the development of these algorithms since processes as randomness, preferential attachment or aging, to name a few, may have crucial implications in the score of the recommendation algorithms. In the current work, we have shown that the traditional paradigm of modeling user-item networks with a combination of preferential attachment and randomness successfully reproduces the degree distribution of both users and items, showing a dominant power-law behavior in movies and stronger exponential dominance for consumers. Nevertheless, this approximation is insufficient for those systems where the interventions of the users report bursting phenomena. To overcome this drawback, we have designed a model of the network growing with explicit temporal correlation in the rating behavior of users. The inclusion of a parameter R, that balances whether the preferential attachment mechanism takes place in the connectivity or in time domain, successfully reproduces several dynamical properties of the network. This is a remarkable result, as the model parameters were obtained just by fitting a static observable, i.e. the node degree distribution. The optimal obtained combination of user node selection strategies shows that 11% of connectivity preferential attachment is enough to fit experimental data when combined with temporal-based prescription. This value is to be compared with the 56% level of connectivity-PA needed when combined with a random selectivity criteria in the QP-model. It is clear then that on one hand, the new introduced mechanism serves to partially grant a "rich-get-richer" scenario regarding connectivity distributions. But most importantly the preferential attachment mechanism in time domain favors the emergence of pattern of burst activity as it favors the succession of consecutive ratings for users who have recently qualified. Moreover the degree-based PA mechanism makes it possible for a user who has not rated for some time to requalify and enter again in a regime of bursts. This selection promotes bringing back users of high degree.

Finally, we think that it is possible to trace an analogy between these results and those obtained in [Zanin *et al.*, 2009], where authors showed that aging effects could increase the efficiency of personal recommendation algorithms. In our case, we believe that the inclusion of the bursting as an *a priori* pattern of users's behavior could also enhance the score of these methods (e.g. to take into account the bursting in order to find the most suitable time to recommend).

Acknowledgments

This work was supported by the Spanish Ministry of S&T [FIS2009-07072] and by the Community of

Madrid under project URJC-CM-2010-CET-5006 and the R&D Program of activities MODELICO-CM [S2009ESP-1691]. It was also supported by University of Buenos Aires through UBACyT grant number 20020090200476 and CONICET by PIP0802/10.

References

Barabási, A.-L. & Albert, R. [1999] "Emergence of scaling in random networks," *Science* **286**, 509–572.

Barabási, A.-L. [2005] "The origin of bursts and heavy tail in human dynamics," *Nature* **435**, 207–211.

Beguerisse Díaz, M., Porter, M. A. & Onnela, J.-P. [2010] "Competition for popularity in bipartite networks," *Chaos* **20**, 043101.

Bennett, J. & Lanning, S. [2007] "The Netflix prize," *KDD Cup and Workshop 2007*, San Jose, California, Aug 12, 2007.

Bobadilla, J., Serradilla, F. & Gutierrez, A. [2009] "Recommender systems: Improving collaborative filtering results," *ICT and Knowledge Engineering*, pp. 100–106.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. [2006] "Complex networks: Structure and dynamics," *Phys. Rep.* **424**, 175–308.

Cano, P., Celma, O., Koppenberger, M. & Buldú, J. M. [2006] "The topology of music recommendation networks," *Chaos* **16**, 013107.

Celma, O. [2010] *Music Recommendation and Discovery*: *The Long Tail*, *Long Fail*, *and Long Play in the Digital Music Space* (Springer-Verlag, Berlin-Heidelberg).

Costa, L. F., Oliveira, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., Viana, M. P. & da Rocha, L. E. C. [2011] "Analyzing and modeling real-world phenomena with complex networks: A survey of applications," *Adv. Phys.* **60**, 329–412.

Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. [2004] "Evaluating collaborative filtering recommender systems," *ACM Trans. Inform. Syst.* **22**, 5–53.

Holme, P., Liljeros, F., Edling, C. R. & Kim, B. J. [2003] "Network bipartivity," *Phys. Rev. E* **68**, 056107.

Netflix on-line stores [2011], http://www.netflix.com.

Newman, M. E. J. [2003] "The structure and function of complex networks," *SIAM Rev.* **45**, 67–256.

Oliveira, J. G. & Barabási, A.-L. [2005] "Darwin and Einstein correspondence patterns," *Nature* **437**, 1251.

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. [2001] "Item-based collaborative filtering recommendation algorithms," *Proc. 10th Int. World Wide Web Conf.* (*WWW10*), May, Hong Kong.

Vazquez, A., Oliveira, J. G., Dezso, Z., Goh, K., Kondor, I. & Barabási, A.-L. [2006] "Modeling bursts and heavy tails in human dynamics," *Phys. Rev. E* **73**, 036127.

Vazquez, A., Rácz, B., Lukács, A. & Barabási, A.-L. [2007] "Impact of non-Poissonian activity patterns on spreading processes," *Phys. Rev. Lett.* **98**, 158702.

Watts, D. J. & Strogatz, S. H. [1998] "Collective dynamics of small-world networks," *Nature* **393**, 440–442.

Zanin, M., Cano, P., Buldú, J. M. & Celma, O. [2009] "Preferential attachment, aging and weights in recommendation systems," *Int. J. Bifurcation and Chaos* **2**, 755–763.

Zhang, Y.-C., Medo, M., Ren, J., Zhou, T., Li, T. & Yang, F. [2007] "Recommendation model based on opinion diffusion," *Europhys. Lett.* **80**, 68003.

Zhou, T., Ren, J., Medo, M. & Zhang, Y. C. [2007] "Bipartite network projection and personal recommendation," *Phys. Rev. E* **76**, 046115.

Zhou, T., Kiet, H. A.-T., Kim, B. J., Wang, B.-H. & Holme, P. [2008] "Role of activity in human dynamics," *Europhys. Lett.* **82**, 28002.

Appendix A

Growth

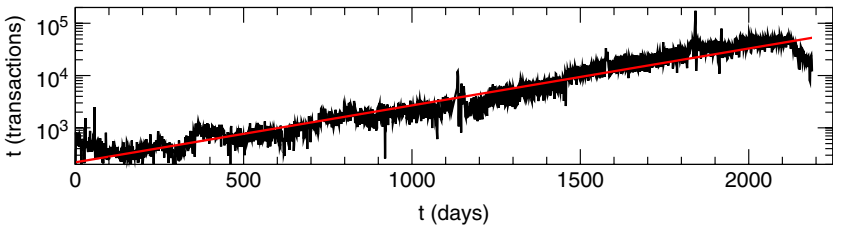
If we take the first rate date as introductory date for a node in the network, we can construct the curves $M(t)$ and $U(t)$ which represent the number of movies and users incorporated to the system at t-days from the beginning. These curves (in black) are shown in panels (b) and (c) of Fig. 7.

In order to incorporate this growing process to the model, we fit both curves in the whole 365 day period following a two-regimes adjustment. For users, this partition is a lineal growing ($1 \leq t \leq 3$) followed by power law curve ($4 \leq t \leq 365$), meanwhile for movies we used a combination of two different power law growths ($1 \leq t \leq 9$) and ($10 \leq t \leq 365$). This scheme resulted in the red curves of Fig. 7.

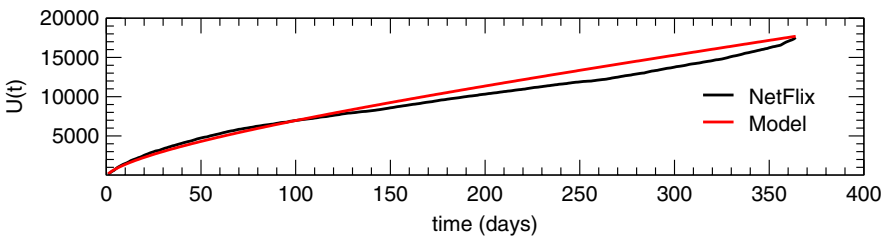
As we said in the description of the QP-model, the time in the original dataset is given in days (labeled as t in this work), meanwhile the natural time scale in the model is in transactions (t_s). The correspondence between days and transactions was fitted from the original dataset following the same approach used in [Beguerisse Díaz *et al.*, 2010] and it is shown in panel (a) of Fig. 7 for the whole Netflix dataset.

Validation of the Extracted One-Year Subset

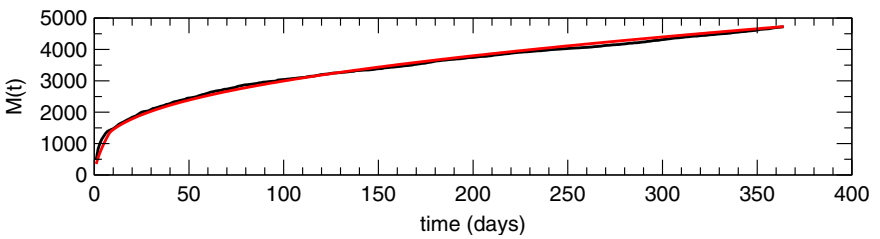
In Fig. 8 we show how the topological properties of bipartite Netflix network for three consecutive years (2001–2003) superimpose each order, validating the



(a)

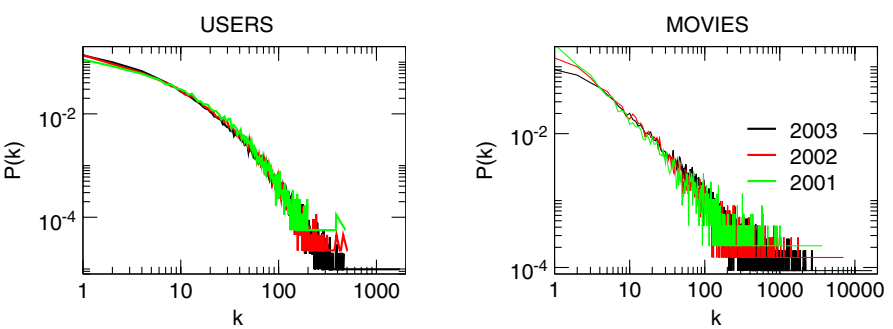


(b)



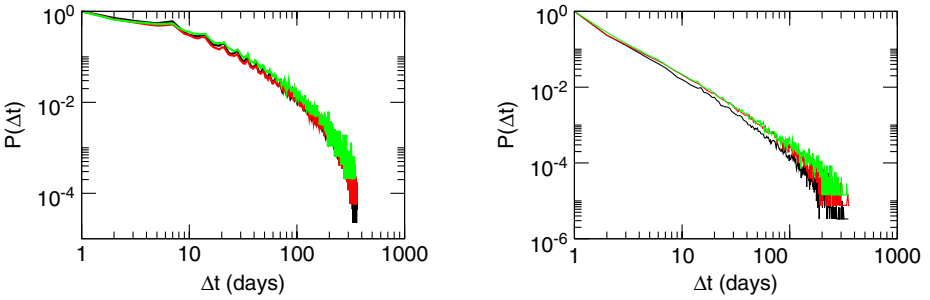
(c)

Fig. 7. Growth process of the network and the model. (a) Number of transactions, (b) users and (c) movies as a function of time. In (a) if for the full 5-year period, (b) and (c) are shown for the 2001 year. Netflix data are in black, meanwhile model curves are displayed in red.



(a) (b)

Fig. 8. Topological properties of bipartite Netflix network for three consecutive years (2001–2003) in one year, temporal windows for (left) users and (right) movies. (a) and (b) Probability of having a degree of k, (c) and (d) normalized distribution of times between transactions, $P(\Delta t)$.



(c) (d)

Fig. 8. (*Continued*)

approach of taking one-year windows to analyze the data.

Model Optimal Fit Parameters

In order to compare the QP and RP models against real data we choose a set of parameters R, P or Q, P which give the best fit in the degree distributions $P(k)$ of users and movies. This adjustment was quantified using the χ^2 of the degree distributions of users and movies for the model when compared with the corresponding real network. Using a step of $\Delta R = \Delta P = \Delta Q = 0.01$ in the region $0.45 < Q < 0.58$ and $0.85 < P < 1.0$ (for the QP model) and $0.04 < R < 0.20$ and $0.9 < P < 1.0$ (for the RP model) we have explored a total of 224 and 187 configurations, respectively.

Since we are interested in finding the best configuration that fits both distributions simultaneously, we minimize the sum of the χ^2 for both types of node. To make the sum and to consider equal weights for both types of nodes, we normalize the value of χ^2 for each distribution (users and movies) according to:

$$\chi^2_{\text{norm}} = \frac{\chi^2 - \min(\chi^2)}{(\max(\chi^2) - \min(\chi^2))},$$

so that the quantity to minimize for the optimal fit will be

$$\chi^2_{\text{NORM}} = \chi^2_{\text{norm(users)}} + \chi^2_{\text{norm(movies)}}. \quad \text{(A.1)}$$

In this way, we find the optimal parameters of the QP and RP models taking the minimum value (A.1) inside the explored region of parameters.