Movies Recommendation Networks as Bipartite Graphs

Jelena Grujić

Scientific Computing Laboratory, Institute of Physics Belgrade,
Pregrevica 118, 11080 Belgrade, Serbia
jelenagr@phy.bg.ac.yu,
http://www.phy.bg.ac.yu

Abstract. In this paper we investigate the users' recommendation networks based on the large data set from the Internet Movie Database. We study networks based on two types of inputs: first (monopartite) generated directly from the recommendation lists on the website, and second (bipartite) generated through the users' habits. Using a threshold number of votes per movie to filter the data, we actually introduce a control parameter, and then by tuning this parameter we study its effect on the network structure. From the detailed analysis of both networks we find that certain robust topological features occur independently from the value of the control parameter. We also present a comparison of the network clustering and shortest paths on the graphs with a randomized network model based on the same data.

Keywords: recommendation networks, bipartite graphs, topology.

1 Introduction

Social networks, representing interactions and relationships between humans or groups, recently became subject of broad research interest [1]. One of the reasons behind this is rapidly evolving electronic technology, which created e-social networks, representing social communication through the Internet. They can be seen as social networks or a technological communication networks [2]. E-mail, chats, web portals, etc., gave us huge amount of information needed for investigated social structures, but also adds new dimension to the social structures. In contrast to the typical communication between pairs of users on the network, such as e-mail network, where a message is sent directly to one known user [3], we can also investigate social structures where users comunicate through common interests, like books, musics, movies, etc. In these user-based dynamical systems subtle correlations between users are developed through hidden feedback mechanisms, in which users share currently available opinions and make actions which, in turn, contribute to further evolution of the network.

Recommendation systems help to overcome information overload by providing personalized suggestions. On many Internet portals, when user selects a product a certain number of alternative products are suggested. These products are nodes of a recommendation network and links point toward the products in their

M. Bubak et al. (Eds.): ICCS 2008, Part II, LNCS 5102, pp. 576–583, 2008.

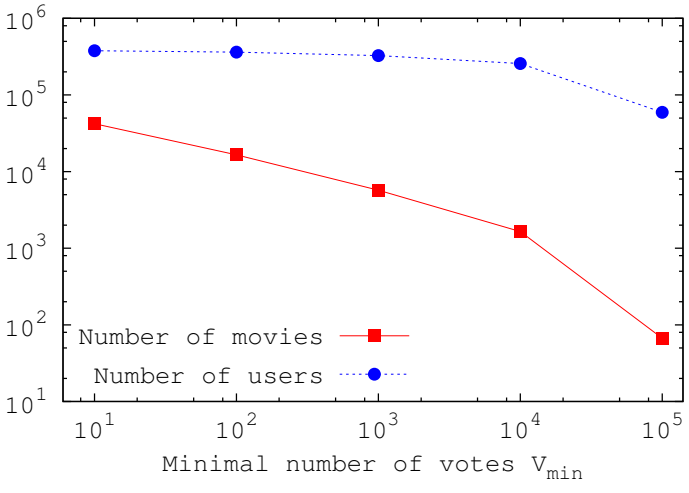

Fig. 1. Dependence of the number of movies and users in the networks on from the control parameter V_{min}, minimal number of votes for a movie in the network

retrospective recommendation lists. For example, networks where products are music groups is studied in Ref. [6]. Recommendations can be generated either through collaborative filtering or using content-based methods or by combination of these two methods. If collaborative filtering is used, the generated network, are actually one-mode projections of bipartite networks, where one type of nodes are products and the other type are users. Links go only between nodes of different types, in this case link is created if the user select the product. Another example connect users with music groups they collected in their music sharing liberties [4,7]. The loss of information in transition from bipartite network to one-mode projection is obvious. In Ref. [5] authors discused the way of obtaining one-mode projection with minimal loss of information and tested their method on the movies-users network.

Here we examined movies-users network using the data from the largest and the most comprehensive on-line movies database IMDb [8]. Generated networks were based on the data on more than $43,000$ movies and $350,000$ users. Furthermore, we introduced a control parameter and tested the universality of our conclusions for the networks of different sizes. We combine two approaches. As a starting point we perform an analysis of the empirical data collected from the website IMDb. Then we investigate the properties of naturally emerging bipartite network based on the users' behavior.

2 Movie Networks

Investigated database IMDb has numerous information regarding more that 1,000,000 films, TV series and shows, direct-to-video product and video games.

The database also posses the user comment, ratings and message boards which characterize users habits. For each movie we collect the following information: ID-number, number of votes V, ID numbers of movies in their recommendation list and ID-numbers of users which commented that movie. Collected information belong to two types of data. First type are IMDb recommendations based on the recommendation system of the site. Second are users' habits, through which, using collaborative filtering, we generate our own recommendation networks. For computational purposes, we concentrated our research only on USA theatrically released movies. This choice was also motivated by the fact that USA theatrical releases had the most comprehensive data.

In order to analyze the universality of our conclusions, we introduced number of votes as control parameter. Users have the opportunity to give their rating for any movie in the Database. Number of votes varies from movie to movie, with more popular ones having more votes. The largest network we analyzed has movies with more that 10 votes and it consists of more that $43,000$ movies and $300,000$ users. We investigated five different sizes of networks according to the minimal number of votes casted for movies $V_{min} \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$. As presumed, number of movies quickly decreases when minimal number of votes increases. V_{min} 1. On the other hand, number of users does not drastically change when V_{min} is increased, which is also expected: if a small number of people voted for some movie than also small number of users commented the movie and by cutting-off huge number of less popular movies you do not cut-off many users. Users usually commented on more than one movie, so by cutting-off some movie you do not necessarily cut-off the user as well. From gathered information we constructed three different networks:

IMDb recommendations (IMDb) monopartite directed network is made by using recommendation system provided by the website [8]. Nodes are movies and links are pointing towards the movies in their respective recommendation list. Rules for generating this recommendation lists is not publicly available (due to the IMDb privacy policy). It is only known that it uses factors such as user votes, genre, title, keywords, and, most importantly, user recommendations themselves to generate an automatic response.

User driven bipartite network (UD-BP) is constructed using users' comments. One type of nodes are movies, and the other type of nodes are users. The movie and the user are linked if a specific user left a comment on the webpage on a specific movie. We do not distinguish between positive and negative comments. As before, we made large bipartite network of almost $43,000$ movies and almost $400,000$ users. Average number of users per movie was 27 with maximum number of $4,763$ comment left for the movie "The Lord of the Rings: The Fellowship of the Ring". Average number of movies per user is much smaller (around three), but the maximal number was $3,040$.

One-mode projection of user driven network (UD-OM) is generated from previous bipartite network. For each movie we generate recommendation list similar to the list provided by the website, but this time based on the users' comments. Like in usual one-mode projection, two movies will be connected if they have

a user which left comment on both movies, but in the recommendation list we put only ten movies with the highest numbers of common users. Among movies with the same number of common users, we choose randomly. We note that this network will also be directed: despite having common users being symmetric relation, if movie i is in the top ten movies of movie j, that does not imply that the movie j is in the top ten movies of movie i.

Users-preferential random network (UP-RD) is generated by connecting each movie to ten other movies randomly from probability distribution proportional to the number of users which left comment on that movie. This way we expect to obtain the degree distribution similar to the distribution of number of users per movie. However since linked movies are chosen randomly, not by their similarity, we expect this to create significant differences in other properties of such network. Test if networks' properties are just the consequence of favoring more popular movies in recommendation system or there is some other underlying mechanism which actually connect movies with movies of their kind.

3 Investigated Properties

In order to determine topology of the network, we focus on the properties which we consider as most importante for the network searchability. Observed properties could lead to possible optimizations of movies recommendation system. Investigated properties are similar to those already studied for real networks:

Degree k_i of the node i is the number of links incident to that node. Average degree $\langle k \rangle$ is k_i averaged over all nodes [1]. The *degree distribution* $P(k)$ is probability that a node selected uniformly at random has degree k. For directed network we can calculate *in*-degree distribution $P(k^{in})$ as distribution of incoming links and *out*-degree distribution $P(k^{out})$ as distribution of outgoing links. Since the number of outgoing links is limited to 10, only degree of incoming links in nontrivial, so we present only those results. In bipartite network we calculate degree distribution for each type of nodes separately.

Clustering coefficient introduced in Ref. [10] expresses how likely is for the two first neighbors j and k of the node i to be connected. Clustering coefficient c_i of the node i is the ratio between the total number e_i of links between its nearest neighbors and the total number of all possible links between these nearest neighbors:

$$c_i = \frac{2e_i}{k_i(k_i - 1).} \tag{1}$$

Clustering coefficient for the whole network $\langle c \rangle$ is average of C_i over all nodes. In directed monopartite networks we did not distinguish between the directions of the links. In bipartite networks there are two types of nodes and links go only between different types, so the above definition for clustering does not apply because triangle does not exist.

A measure of the typical separation between two nodes in the graph is given by the *average shortest path length* D, defined as the mean of all shortest paths lengths d_{ij} [10]. A problem with this definition is that D diverges if network

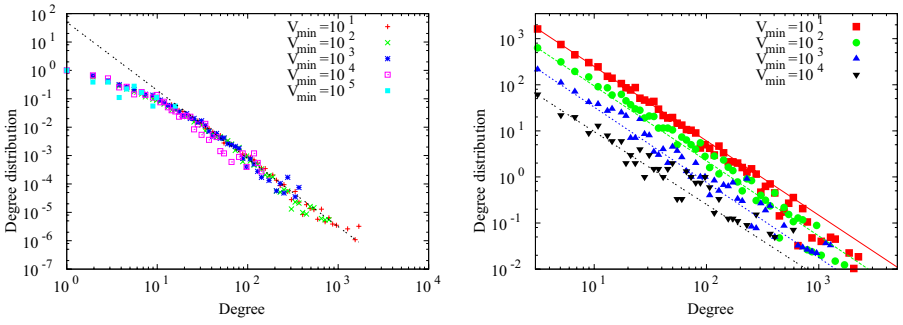

Fig. 2. Degree distributions for networks with different minimal number of votes V_{min}. On the left directed monopartite networks based on IMDb recommendations. The tail is fitted with the power-law $k^{-1.8}$. On the right for user driven one-mode projection (UD-OM) fitted to the power law $k^{-1.6}$.

is not connected, as in our example. Here we calculate D as the average of all existing shortest paths. We also use alternative approach and considered the harmonic mean [11] of the shortest paths, so-called *topological efficiency*.

$$D = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \qquad E = \frac{1}{N(N-1)} \sum_{i,j,i \neq j} \frac{1}{d_{ij}}. \tag{2}$$

4 Results and Discusion

A common feature which appears in all studied networks is a degree distribution with power law tail. Even more important is the fact that the degree distributions are universal for networks of different sizes, obtained by changing the control parameter. This suggests that even though we studied large but limited number of movies, our main result do not depend on the number of movies, i.e. on finite size effect.

For user driven bipartite network, the distribution of number of movies per user is very robust power-law, universal for all sizes of networks. The distribution fits to the power law with the exponent 2.16. This exponet occurs in most of the studied real world networks [1]. The distribution of number of users per movie (Fig 3) is well described by a power law for the largest investigated networks ($V_{min} > 10$), for the smaller networks this is not the case, as the number of movies per user decreases below 20. This is expected since in smaller networks we do not have less popular movies, and those are the movies which usually have small number of users.

Even though IMDb recommendations network has degree distributions which are not power laws for the degrees less than 11, the distributions can be rescaled so as to fit to the same curve even for the small degrees. Like their bipartite

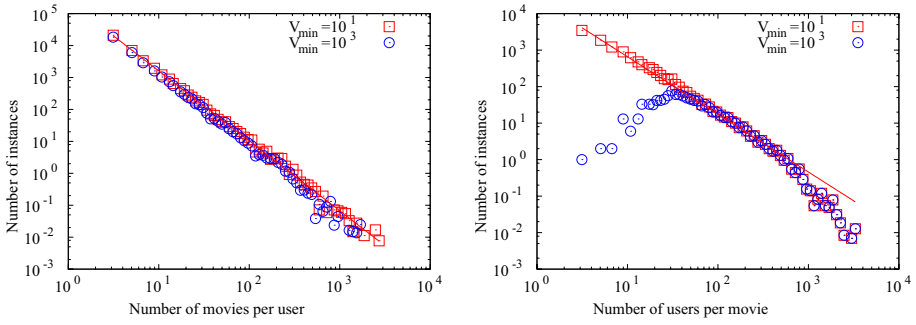

Fig. 3. Degree distributions for bipartite networks for $V_{min} = 10^1$ and $V_{min} = 10^3$. On the left graph the number of movies per user fit $k_u^{-2.19}$, while on the right graph the number of users per movie fit $k_m^{-1.58}$. Distributions are logarithmically binned to reduce the noise.

counterparts, one-mode projections have the power-law degree distributions through the whole range of degrees (Fig 2). Exponent of the power law is close to the exponent of the distribution of the number of users per movie. We emphasize that we did not performe separate one-mode projection of network of different sizes. Rather, we constructed one-mode projection for the largest network and then constructed smaller networks by eliminating movies with less that V_{min} votes and all of their links from the largest network. All distributions are logarithmically binned in order to decrease the noise.

Both networks based on the real data show small world property. Clustering coefficients are high and are increasing when size of the network decreases. Since smaller networks are missing less popular movies, more popular movies networks are more clustered. Average path lengths are small and decreasing with the size of the network. Topological efficiency is increasing for smaller networks (Fig 4).

As expected, the degree distribution of the users-preferential random network is a power-law with the similar exponent as IMDb and UD-OM networks. But apart from the degree distribution, other properties of are significantly different. Most obvious difference is in the clustering coefficient. As expected, random networks have the clustering coefficient few orders of magnitude smaller than those of the real networks. Average shorthest paths are significantly smaller although they exhibit similar behavior. Also, we see that the efficiencies are proportionally larger. We note that some properties of IMDb network are closer to the ones of UP-RD networks. Power-law degree distribution of UD-OM could be a consequence of the preferential attachment to more popular movies. During construction of the network we connected movies with more common users. Movies with more users would also have greater probability to have more common users with some other movie. However, we see that if we connect movies only by favoring more popular movies, other properties would be different. Smallest network with $V_{min} = 10^5$ is so sparse that we eliminated it from the investigation.

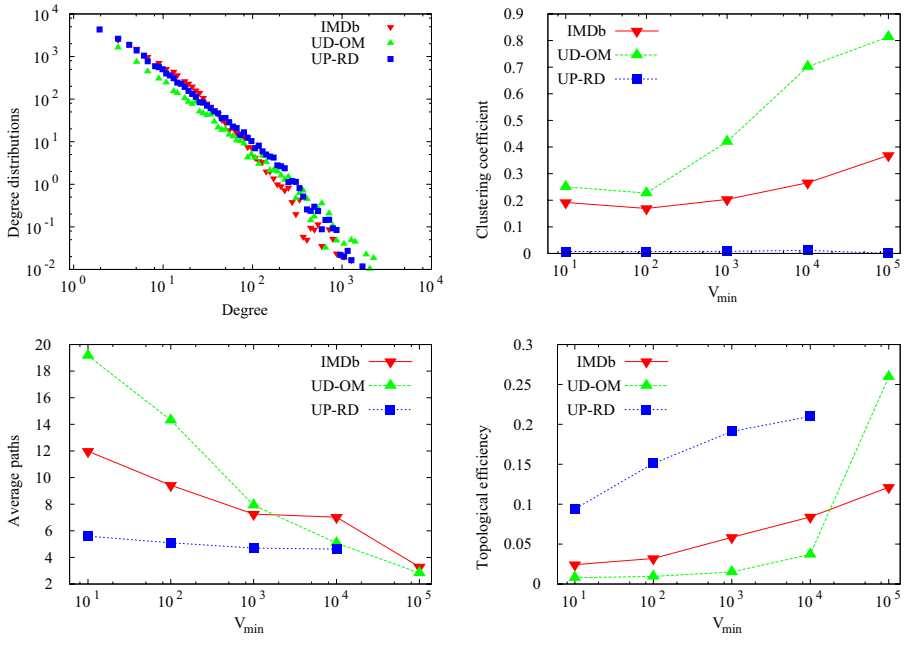

Fig. 4. Compareson of IMDb recommendations (IMDb), User driven monopartite directed (UD-OM) and Users-Preferentail random (UP-RD) networks. Degree distribution for $V_{min} = 10$ (top left), clustering coefficient (top right), average shortest paths lengths (bottom left), topological efficiency (bottom right) as a function of V_{min}.

5 Conclusion and Future Directions

Using the data from the largest and most comprehensive movie database IMDb, we considered two types of networks: one based on the IMDb recommendations and one based on collaborative filtering based on user habits. As a starting point we investigated properties of movies directed network following directly from IMDb recommendations data. We generated bipartite networks by connecting users with the movies they commented on. In order to compare these two approaches, we made one-mode projection of bipartite networks. We introduced the minimal number of votes as a control parameter and constructed different sizes of networks according to the number of votes of movies. All networks show high clustering coefficients and small world property, although some variations are noticed in the behavior for different sizes of networks. Degree distributions for both types of networks are universal. Networks obtained through collaborative filtering exhibits robust power-low distributions seemingly universal for all sizes of networks. Networks based on IMDb recommendations although not power law distributions for small values of degrees, still have power-law tail. Since the properties of random vote-preferential networks, most noticeably clustering coefficient are significantly different from one-mode projection, we believe

that the power-low distribution is not just the consequence of the favoring more popular movies, but some self-organizing mechanism.

In the future, we plan to generalize the presented approach. By investigating the community structures, users clustering and by further theoretical modeling, we are going to try to understand natural mechanisms behind these properties.

Acknowledgments. Author thanks B. Tadić for numerous useful comments and suggestions. This cooperation is supported by COST-STSM-P10-02988 and PATTERNS project MRTN-CT-2004-005728. This work was financed in part by the Ministry of Science of the Republic of Serbia under project no. OI141035. and FP6 project CX-CMCS. The presented numerical results were obtained on the AEGIS GRID e-infrastructure whose operation is supported in part by FP6 projects EGEE-II and SEE-GRID-2.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Huang, D.-U.: Complex Networks: Structure and dynamics. Phys. Rep. 424, 175–308 (2006)
2. Tadić, B., Rodgers, G.J., Thurner, S.: Transpot on complex networks: flow, jamming and optimization. Int. J. Bifurcation and Chaos (IJBC) 17(7), 2363–2385 (2007)
3. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Phys. Rev. E 68, 065103 (2003)
4. Lambiotte, R., Ausloos, M.: Uncovering collective listening habits and music genres in bipartite networks. Phys. Rev. E 72, 066107 (2005)
5. Zhou, T., Ren, J., Medo, M., Zhang, Y.C.: Bipartite network projection and personal recommendation. Phys. Rev. E 76, 046115 (2007)
6. Cano, P., Celma, O., Koppenberger, M.: Topology of music recommendation networks. Chaos 16, 013107 (2006)
7. Lambiotte, R., Ausloos, M.: On the genre-fication of Music: a percolation approach. Eur. Phys. J. B 50, 183–188 (2006)
8. Internet Movie Database, `http://www.imdb.com`
9. Lind, P., González, M., Herrmann, H.: Cycles and clustering in bipartite networks. Phys. Rev. E 72, 056127 (2005)
10. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393, 440–442 (1998)
11. Marchiori, M., Latora, V.: Harmony in the Small-World. Physica A 285, 198701 (2000)