COMPLEX NETWORKS – MAI

Film recommender

www.filmaffinity.com

José A. Magaña Mesa

46703957L

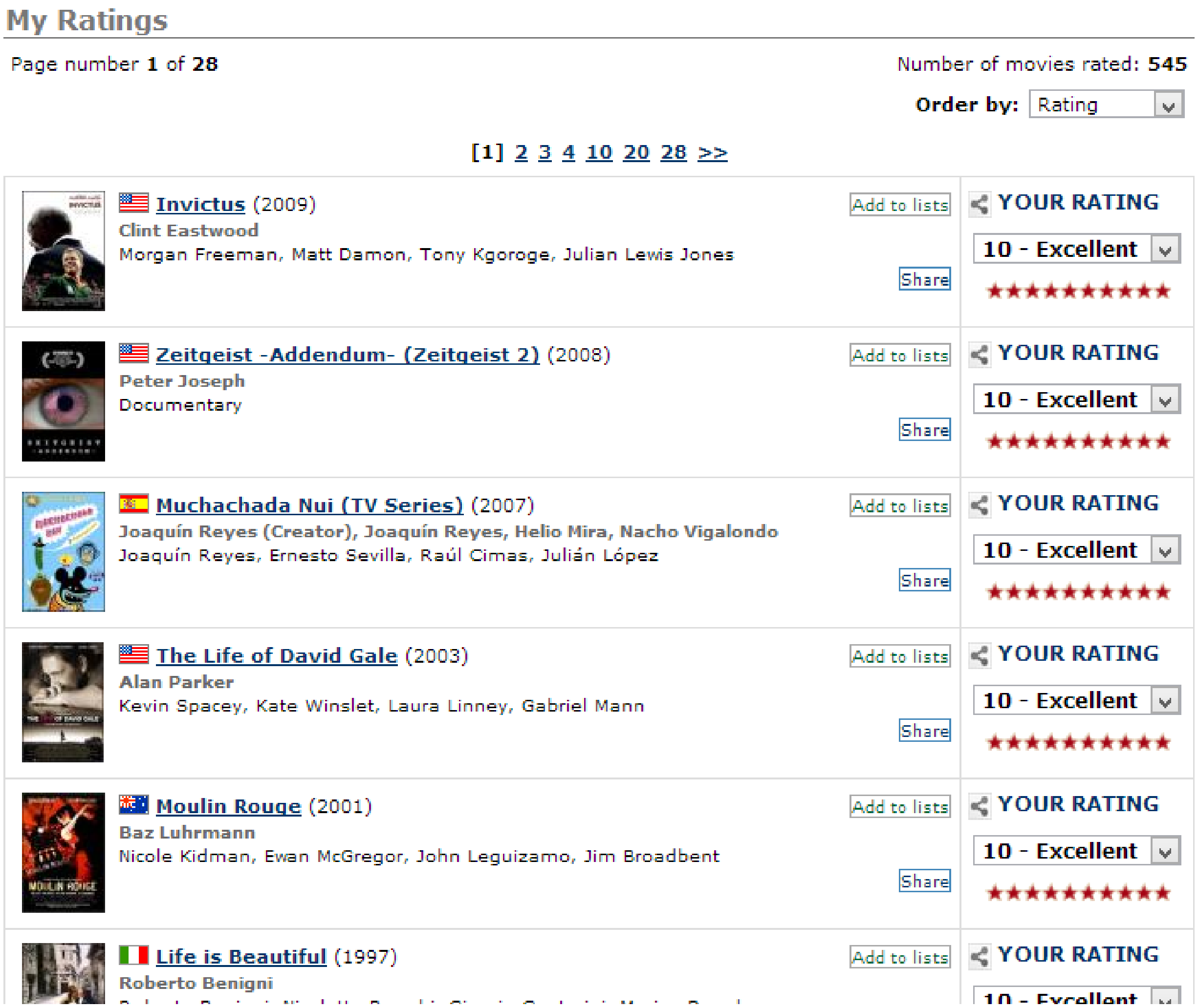
June 25th, 2013

FILM AFFINITY

How it works

Film Affinity (www.filmaffinity.com) is a website that recommends movies to its users.

For doing so, the web ask the user to score from 0 to 10 the movies the user has already watched.



My Ratings

Page number **1** of **28**

Number of movies rated: **545**

Order by: Rating

[1] 2 3 4 10 20 28 >>

Invictus (2009)
Clint Eastwood
Morgan Freeman, Matt Damon, Tony Kgoroge, Julian Lewis Jones
Add to lists | Share
YOUR RATING: 10 - Excellent ★★★★★★★★★★

Zeitgeist -Addendum- (Zeitgeist 2) (2008)
Peter Joseph
Documentary
Add to lists | Share
YOUR RATING: 10 - Excellent ★★★★★★★★★★

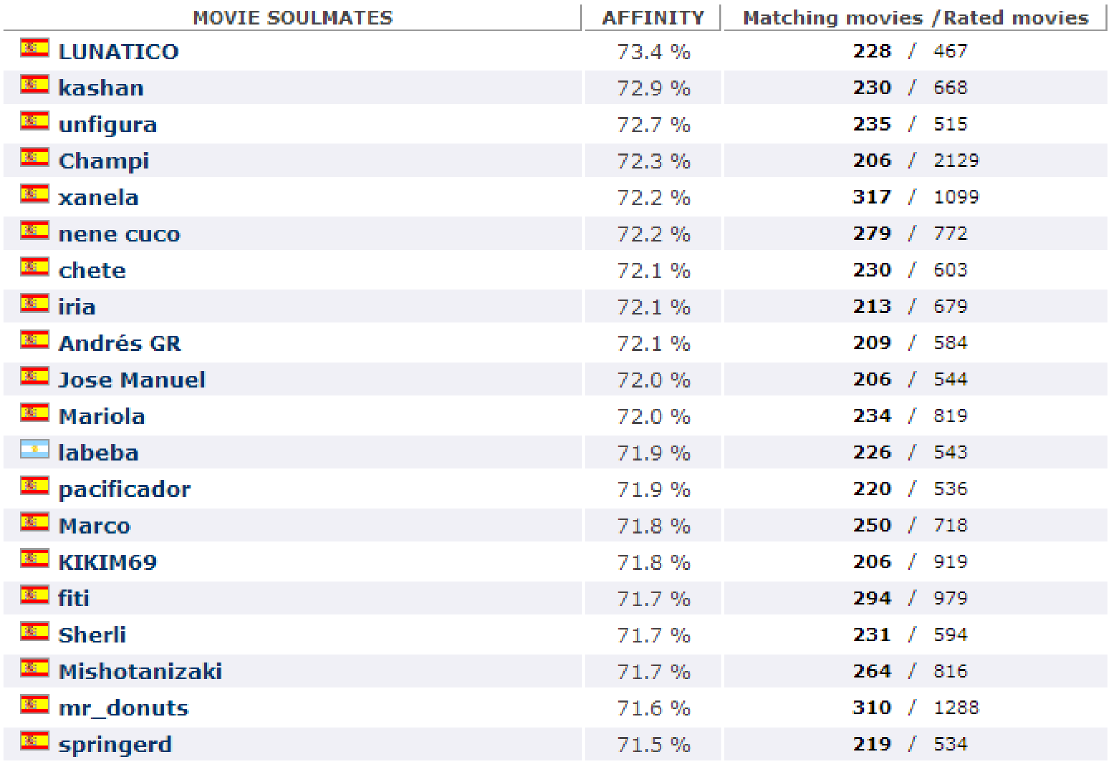
Muchachada Nui (TV Series) (2007)
Joaquín Reyes (Creator), Joaquín Reyes, Helio Mira, Nacho Vigalondo
Joaquín Reyes, Ernesto Sevilla, Raúl Cimas, Julián López
Add to lists | Share
YOUR RATING: 10 - Excellent ★★★★★★★★★★

The Life of David Gale (2003)
Alan Parker
Kevin Spacey, Kate Winslet, Laura Linney, Gabriel Mann
Add to lists | Share
YOUR RATING: 10 - Excellent ★★★★★★★★★★

Moulin Rouge (2001)
Baz Luhrmann
Nicole Kidman, Ewan McGregor, John Leguizamo, Jim Broadbent
Add to lists | Share
YOUR RATING: 10 - Excellent ★★★★★★★★★★

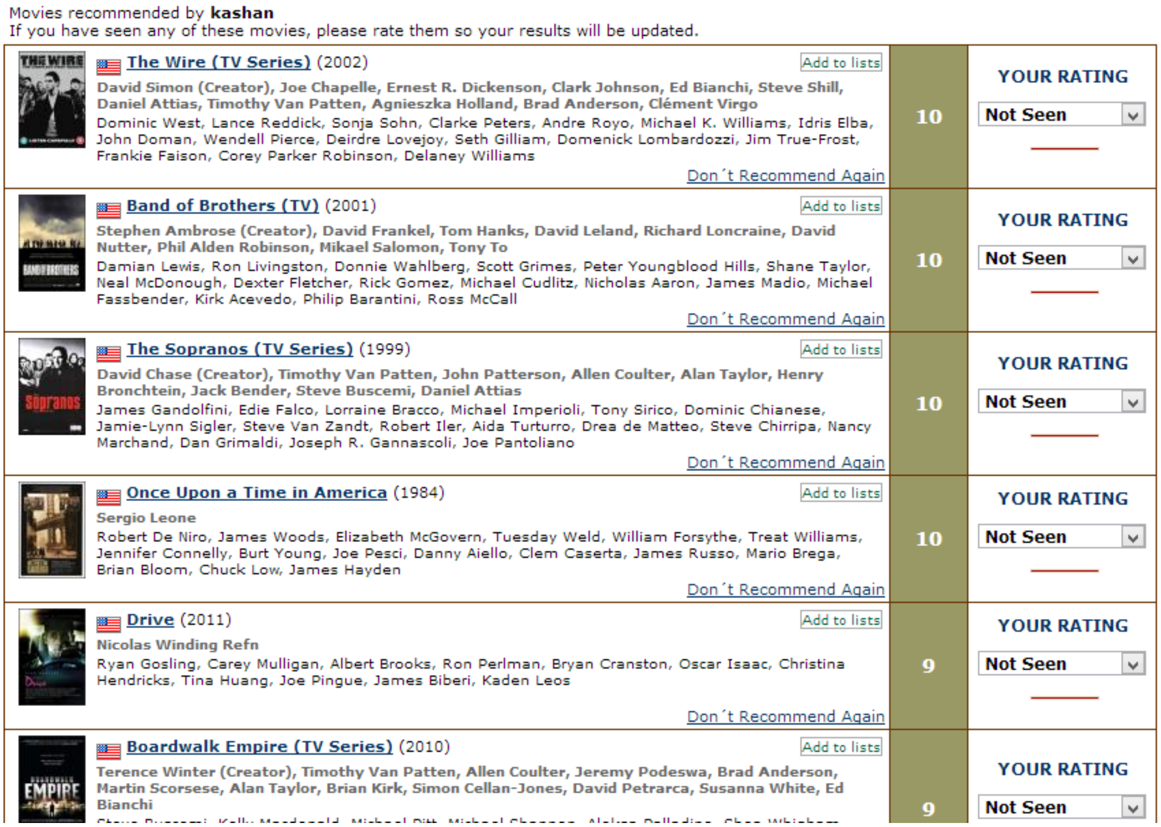
Life is Beautiful (1997)
Roberto Benigni
Add to lists
YOUR RATING: 10 - Excellent

Based on that, and the information that other users have also provided (scoring their movies), the web selects the 20 users that have the closest scorings to the own user (known as movie soulmates). For each one of these users, an affinity ratio is calculated and the users are sorted according to it. This affinity is calculated based on how many movies (in absolute number) both you and your soulmate have watched and how close have been your scorings for these movies.



MOVIE SOULMATES	AFFINITY	Matching movies / Rated movies
🇪🇸 LUNATICO	73.4 %	**228** / 467
🇪🇸 kashan	72.9 %	**230** / 668
🇪🇸 unfigura	72.7 %	**235** / 515
🇪🇸 Champi	72.3 %	**206** / 2129
🇪🇸 xanela	72.2 %	**317** / 1099
🇪🇸 nene cuco	72.2 %	**279** / 772
🇪🇸 chete	72.1 %	**230** / 603
🇪🇸 iria	72.1 %	**213** / 679
🇪🇸 Andrés GR	72.1 %	**209** / 584
🇪🇸 Jose Manuel	72.0 %	**206** / 544
🇪🇸 Mariola	72.0 %	**234** / 819
🇦🇷 labeba	71.9 %	**226** / 543
🇪🇸 pacificador	71.9 %	**220** / 536
🇪🇸 Marco	71.8 %	**250** / 718
🇪🇸 KIKIM69	71.8 %	**206** / 919
🇪🇸 fiti	71.7 %	**294** / 979
🇪🇸 Sherli	71.7 %	**231** / 594
🇪🇸 Mishotanizaki	71.7 %	**264** / 816
🇪🇸 mr_donuts	71.6 %	**310** / 1288
🇪🇸 springerd	71.5 %	**219** / 534

Then for each of your 20 soulmates, a list of 25 movies is suggested to you. This list of 25 movies is not filtered by any other criteria than the scoring of the soulmate.



Movies recommended by **kashan**
If you have seen any of these movies, please rate them so your results will be updated.

🇺🇸 **The Wire (TV Series)** (2002) — 10
David Simon (Creator), Joe Chapelle, Ernest R. Dickenson, Clark Johnson, Ed Bianchi, Steve Shill, Daniel Attias, Timothy Van Patten, Agnieszka Holland, Brad Anderson, Clément Virgo
Dominic West, Lance Reddick, Sonja Sohn, Clarke Peters, Andre Royo, Michael K. Williams, Idris Elba, John Doman, Wendell Pierce, Deirdre Lovejoy, Seth Gilliam, Domenick Lombardozzi, Jim True-Frost, Frankie Faison, Corey Parker Robinson, Delaney Williams

🇺🇸 **Band of Brothers (TV)** (2001) — 10
Stephen Ambrose (Creator), David Frankel, Tom Hanks, David Leland, Richard Loncraine, David Nutter, Phil Alden Robinson, Mikael Salomon, Tony To
Damian Lewis, Ron Livingston, Donnie Wahlberg, Scott Grimes, Peter Youngblood Hills, Shane Taylor, Neal McDonough, Dexter Fletcher, Rick Gomez, Michael Cudlitz, Nicholas Aaron, James Madio, Michael Fassbender, Kirk Acevedo, Philip Barantini, Ross McCall

🇺🇸 **The Sopranos (TV Series)** (1999) — 10
David Chase (Creator), Timothy Van Patten, John Patterson, Allen Coulter, Alan Taylor, Henry Bronchtein, Jack Bender, Steve Buscemi, Daniel Attias
James Gandolfini, Edie Falco, Lorraine Bracco, Michael Imperioli, Tony Sirico, Dominic Chianese, Jamie-Lynn Sigler, Steve Van Zandt, Robert Iler, Aida Turturro, Drea de Matteo, Steve Chirripa, Nancy Marchand, Dan Grimaldi, Joseph R. Gannascoli, Joe Pantoliano

🇺🇸 **Once Upon a Time in America** (1984) — 10
Sergio Leone
Robert De Niro, James Woods, Elizabeth McGovern, Tuesday Weld, William Forsythe, Treat Williams, Jennifer Connelly, Burt Young, Joe Pesci, Danny Aiello, Clem Caserta, James Russo, Mario Brega, Brian Bloom, Chuck Low, James Hayden

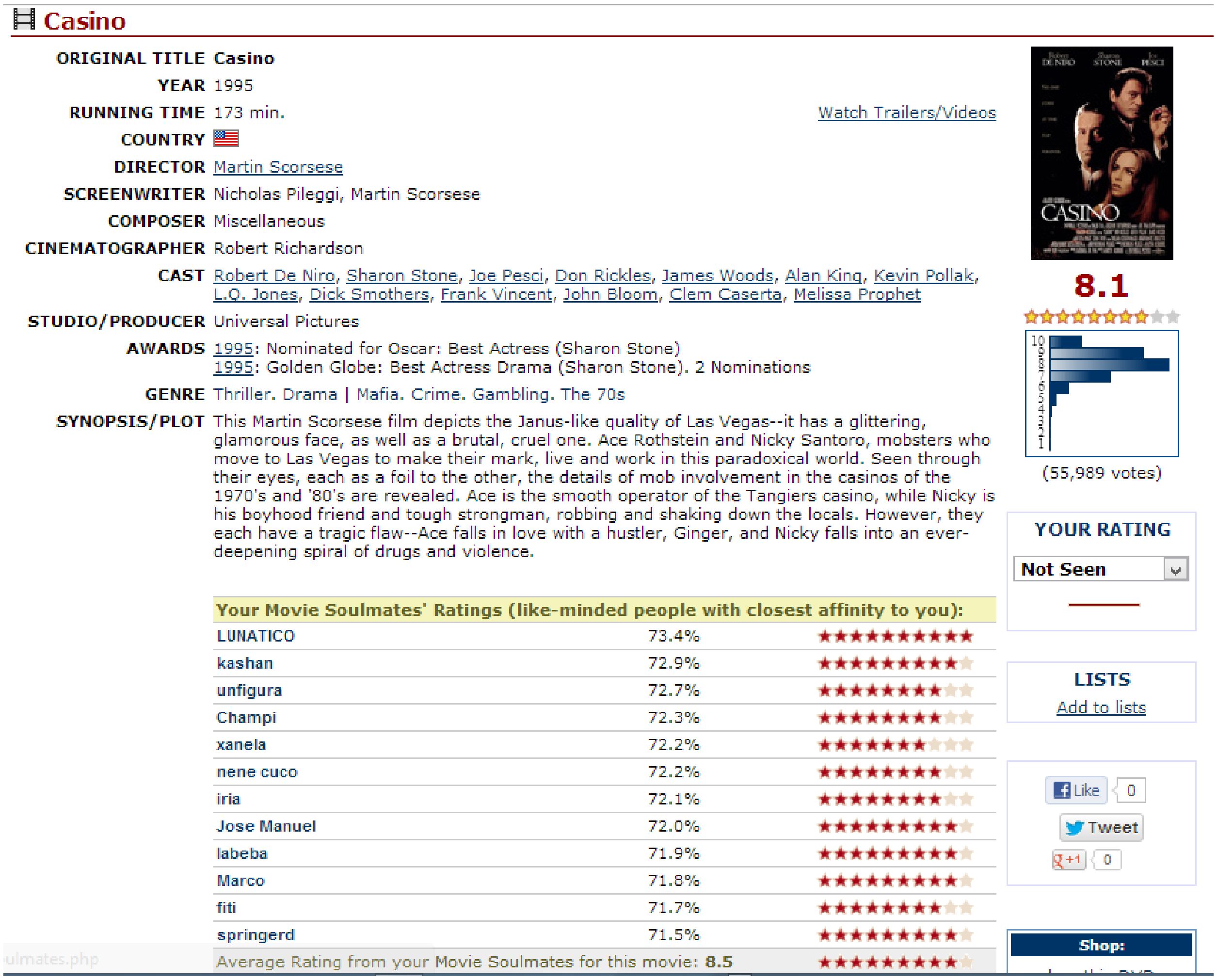
🇺🇸 **Drive** (2011) — 9
Nicolas Winding Refn
Ryan Gosling, Carey Mulligan, Albert Brooks, Ron Perlman, Bryan Cranston, Oscar Isaac, Christina Hendricks, Tina Huang, Joe Pingue, James Biberi, Kaden Leos

🇺🇸 **Boardwalk Empire (TV Series)** (2010) — 9
Terence Winter (Creator), Timothy Van Patten, Allen Coulter, Jeremy Podeswa, Brad Anderson, Martin Scorsese, Alan Taylor, Brian Kirk, Simon Cellan-Jones, David Petrarca, Susanna White, Ed Bianchi
Steve Buscemi, Kelly Macdonald, Michael Pitt, Michael Shannon, Aleksa Palladino, Shea Whigham

Available information

The web provides access to your own list of scored movies and also the whole list of your soulmates (apart from the movies that your soulmates recommends to you).

For every movie a complete description of the movie is provided (actors, director, screenwriter, genre, …). Only at the movie information, the scoring that all your soulmates have assigned is shown and an average score is provided (but not weighted based on the affinity of your soulmates).



Casino

ORIGINAL TITLE Casino
YEAR 1995
RUNNING TIME 173 min.
COUNTRY
DIRECTOR Martin Scorsese
SCREENWRITER Nicholas Pileggi, Martin Scorsese
COMPOSER Miscellaneous
CINEMATOGRAPHER Robert Richardson
CAST Robert De Niro, Sharon Stone, Joe Pesci, Don Rickles, James Woods, Alan King, Kevin Pollak, L.Q. Jones, Dick Smothers, Frank Vincent, John Bloom, Clem Caserta, Melissa Prophet
STUDIO/PRODUCER Universal Pictures
AWARDS 1995: Nominated for Oscar: Best Actress (Sharon Stone)
1995: Golden Globe: Best Actress Drama (Sharon Stone). 2 Nominations
GENRE Thriller. Drama | Mafia. Crime. Gambling. The 70s
SYNOPSIS/PLOT This Martin Scorsese film depicts the Janus-like quality of Las Vegas--it has a glittering, glamorous face, as well as a brutal, cruel one. Ace Rothstein and Nicky Santoro, mobsters who move to Las Vegas to make their mark, live and work in this paradoxical world. Seen through their eyes, each as a foil to the other, the details of mob involvement in the casinos of the 1970's and '80's are revealed. Ace is the smooth operator of the Tangiers casino, while Nicky is his boyhood friend and tough strongman, robbing and shaking down the locals. However, they each have a tragic flaw--Ace falls in love with a hustler, Ginger, and Nicky falls into an ever-deepening spiral of drugs and violence.

Watch Trailers/Videos

8.1

(55,989 votes)

YOUR RATING

Not Seen

LISTS

Add to lists

Your Movie Soulmates' Ratings (like-minded people with closest affinity to you):		
LUNATICO	73.4%	★★★★★★★★★★
kashan	72.9%	★★★★★★★★★☆
unfigura	72.7%	★★★★★★★★☆☆
Champi	72.3%	★★★★★★★★☆☆
xanela	72.2%	★★★★★★★☆☆☆
nene cuco	72.2%	★★★★★★★★☆☆
iria	72.1%	★★★★★★★★☆☆
Jose Manuel	72.0%	★★★★★★★★★☆
labeba	71.9%	★★★★★★★★★☆
Marco	71.8%	★★★★★★★★★☆
fiti	71.7%	★★★★★★★★☆☆
springerd	71.5%	★★★★★★★★★☆
Average Rating from your Movie Soulmates for this movie: **8.5**		★★★★★★★★★☆

Limitations

A single list is not provided based on any criteria that conjugate the 20 lists (one for every soulmate) that are suggested to the user what would be a much valuable recommendation as it would consider the opinion of all your soulmates.

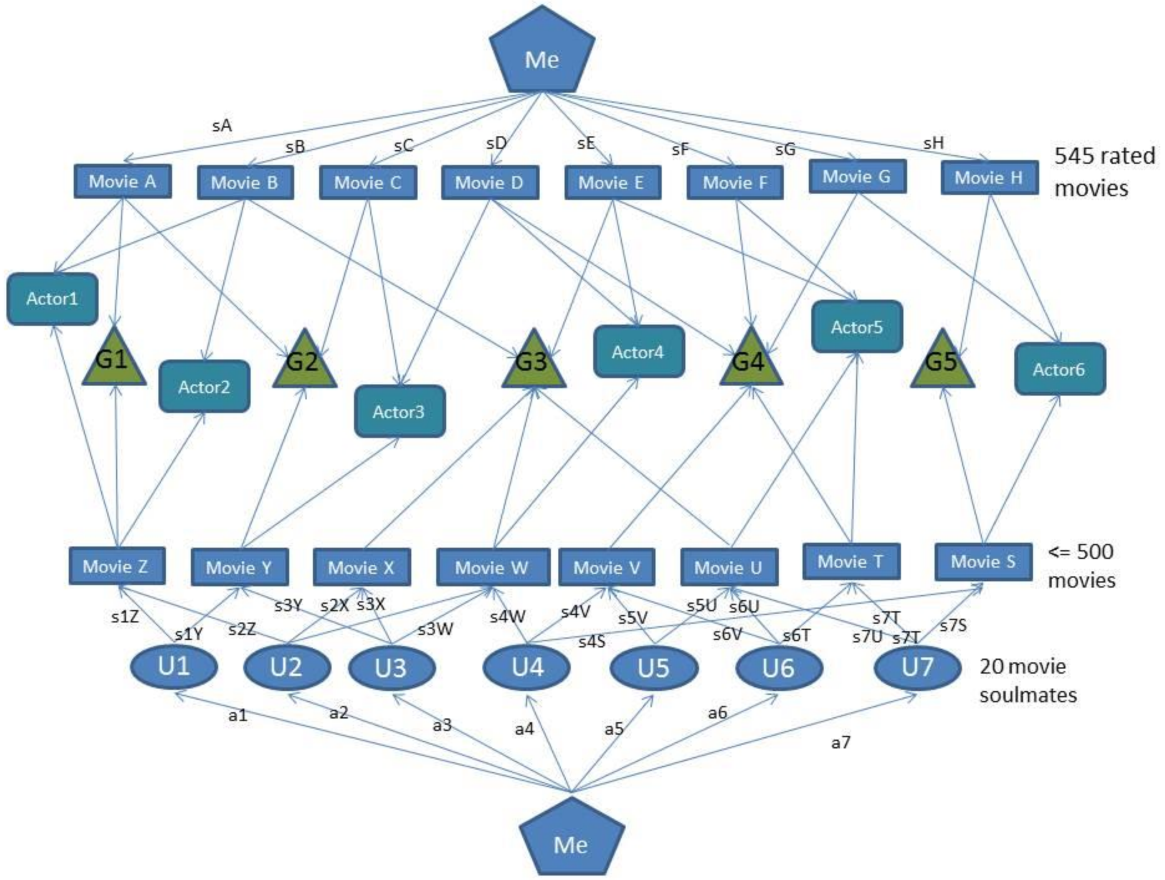
The information conjugating all the scores can only be analyzed, browsing movie by movie. It is then difficult to know the BEST recommendation.

Proposal

Create a network that conjugates all the information for the user, that is:

- The list of movies scored by the user (546 in my case)
- The list of soulmates, that by using the different criterion that the website provides to select their characteristics (age, country of origin, match), that contains 274 users.
- For every movie (either on the user list or the soulmate recommendation list) consider some of the information provided, e.g: actors and genre.

This constitutes a network that is a multi-partite graph as the network will have different type of nodes (Movies, Users (soulmates), Actors, Genre).



Based on this network, and given the limitations found during data capture, the objectives have been:

- Analyze the network defined by movies, users and movie rates
- Verify how the use of social information helps to improve the estimation of my preferences compared to use simply the average rate of the movie comparing to the real rate provided by the reference user (me).
- Use the obtained method to consolidate a list of recommended movies (not already watched).

For answering these questions, and avoid falling in solutions in other disciplines, the algorithm will search to rate solutions based on node weighted degrees (**frequency**).

Intuitively, and this can researched during the project, the network has a **preferential attachment** (popular movies become more popular) bias but it is likely also to observe a **small world** effect (movies not similar to the ones watched by a user are recommended). Users with a high number of scored movies will become strong influencers (**opinion leads**), as they have a higher chance of having a high percentage of matches in the list of watched movies with other users (in the way currently calculated).

This is especially true considering that the Film Affinity recommender does not consider defining the affinity with a user based on the movies that your candidate soulmate has seen but you have not.

Data collection

In order to collect the data from the FilmAffinity website, a crawler has been used. The crawler used is weblech 0.0.3 that even being an Alpha version not evolved since 2004, has allowed with limited modifications fulfill the necessary data extraction.

The crawler has been modified in order to selectively navigate only the subset of links that contain the information required for the experiment: users, movies, ratings. All other information, like images or links to other information not required has been removed from the crawling list.

A second modification introduced has been the time among requests. This modification has prevented the crawling from being detected by allowing 1 second between consecutive gets at the expense of slowing down the process.

For the cases in which the crawling has been interrupted, a third modification has allowed to initialize the list of URLs with the URL's in a set of files already downloaded. This functionality has also been useful to bring the crawler beyond some points that required user interaction. Some pages have been manually downloaded and then used as seed to launch the crawler again.

Also to overcome the authentication on some parts of the website, manual download of some pages have been required as the authentication implemented by the website does not allow preemptive authentication.

Regardless all these measures during the early phases of the data extraction the website has banned our IP twice.

From the data extraction, the following information has been collected

	Rated	**Unrated (0 VOTES)**	**User rated**
Films	43678	27069	19777
My rates	546		
			User rates
Users	274		187320

Generated networks

From the extracted data, the following networks are built directly:

Movie average ratings

Every movie has an average rate based on the rates given by all users who have rated the movie, also the number of votes received is known, even when the individual rating given by every user is not provided. This information is indeed just half of the total network but from it, it is possible to infer the distribution of votes for the whole network.

This will be used to analyze the configuration of the network and validate the hypothesis found in papers about the expected degree distribution.

This network accounts for 70747 movies, of which only 43678 have an average rate. It must be said that the website only provides an average rate for movies that have more than 10 votes. For movies with less than 10 votes, no average rate is provided.

User movie ratings

The main network obtained consists on the individual rates given to movies by each of the users identified by the web as my soulmates.

A total of 274 users have provided 187320 rates to 19777 distinct movies. The network will have then, 20051 nodes, corresponding to the sum of users and movies and 187320 edges.

Some measurements of the network, extracted by Gephi, can be found on section

Appendix: Gephi user-movie rates. Some comments to the results follow.

The network is connected with diameter equal to 4 meaning that from any user there will always be someone who has seen at least one movie that we have seen that has seen any other movie on the list. The eccentricity, 3, confirms this.

In the case of movies, where in all cases eccentricity is 4, any movie is connected to any other movie in the list by a path that goes through two users.

Because of the construction of the network, the node with higher betweenness corresponds to my user.

If the betweenness of movies is analyzed and we obtain that the top movies correspond to great hits:

Movie	Betweennes	Title
m444796	510849.0769	Forrest Gump
m160882	468135.1254	Pulp Fiction
m230028	449921.5122	Slumdog Millionaire
m768790	443755.918	The Silence of the Lambs
m594480	442262.8869	Life is Beautiful
m161026	441365.7565	The Shawshank Redemption
m656153	436051.4407	Schindler's List
m314359	432757.7419	Million Dollar Baby
m539054	428822.0273	Gran Torino
m575149	427616.1737	Seven (Se7en)
m505307	418075.8004	American Beauty
m392075	411121.4022	Gladiator
m809297	410172.3868	The Godfather
m745383	403790.9315	A Clockwork Orange
m746997	402860.4813	Inglourious Basterds
m814379	399851.6586	Titanic
m607127	369189.0233	The Sixth Sense
m360471	366568.5769	The Curious Case of Benjamin Button
m598422	363871.8619	The Shining
m121917	362887.9906	Mystic River

The clustering of the graph produces the following classes:

Class ID	Users	Movies	Normalized Ratio Movie-User
0	8	1726	215.75
1	4	435	108.75
2	27	2646	98.00
3	26	3405	130.96
4	57	7677	134.68
5	144	2701	18.76
6	9	1187	131.89
	275	**19777**	71.92

It is relevant that the separation of users is so uneven and for example class 0 has with only 8 users, such a great number of movies in the class. This parameter looks important to be analyzed as a characteristic of bipartite networks. This study goes beyond the objective of this work.

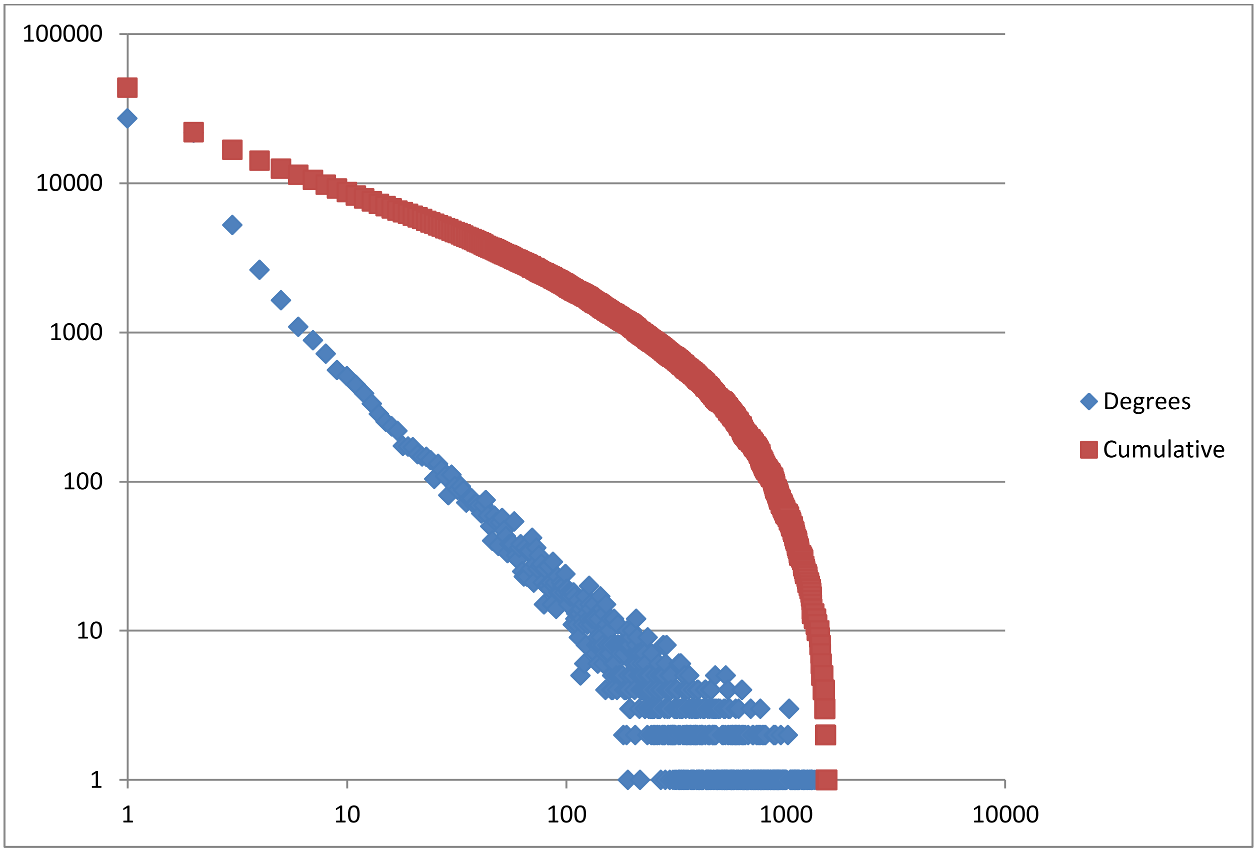
All the movies shown in the rank of betweenness belong to class 5. In fact in the top 100 betweenness movies, 90 nodes belong to class 5. This goes in the direction of showing that very popular movies form very dense subnetworks with a high number of users. Nevertheless, the modularity obtained is 0.220 indicating that the clusters obtained are not very compact, and could be easily redistributed.

Preferential attachment

It is stated in [2] that networks that are created by users, in contrast to networks created by experts, follow a preferential attachment creation scheme and hence, the degree distribution follows a Power Law.

This has been checked using the complete list of movies (70747) and the number of votes that all users in the system have provided.

The result can be seen, in log-log scale in the following graph, for both the degree distribution and the accumulated degree distribution. The degrees have been grouped in steps of 100.

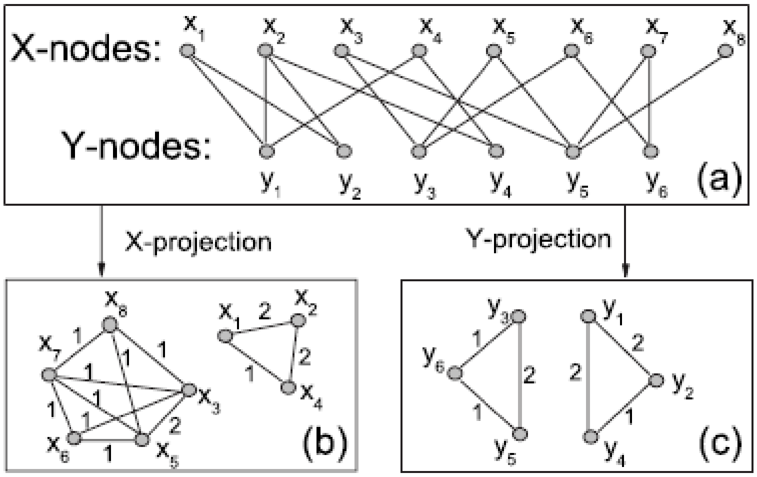


The graph shows the differential features of a Power Law distribution, with a long tail of nodes with low degree and a linear decay (in log-log scale).

Graph transformation: projection

Bipartite projection

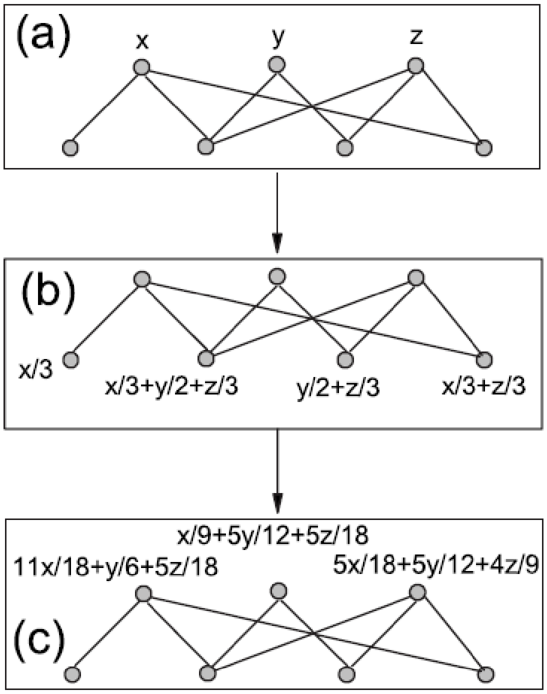
The original network is a bipartite network, with movies being one part of the network and users the other part. The edges represent the rating for a movie given by the user. This factor is an integer between 0 and 10. In order to further analyze the network it is required to separate the nodes in the two types of nodes, movies and users. An example of a simple way of doing this separation is shown:


X-nodes:
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Y-nodes:
y_1 y_2 y_3 y_4 y_5 y_6
(a)
X-projection
Y-projection
x_8 x_7 x_3 x_6 x_5 x_1 x_2 x_4
1 1 1 1 1 1 1 1 2 1 2 2 1
(b)
y_3 y_6 y_5 y_1 y_2 y_4
1 2 1 2 2 1
(c)

In this example, the number of commons nodes is used to weight the created synthetic node.

The idea of projecting the original bipartite graph for each of its two parts, has been extracted from [1], that shows a binary version, where users have closer similarity when both of them have acquired is the same product regardless the rating given to the product. The cosine similarity is then used.

A second study where the projection of a bipartite network is calculated is [11]. In this case, the weight of the node is calculated based on the linear combination of the common nodes, considering the neighbor output degree. The next figure illustrates the process:



In this case, the node requires having an attribute whose value is used to calculate the projected edge weight. In our case, both, movies and users have an average rate that could be used but it has been considered, and verified with some calculations, that the weight would be very biased towards nodes with higher degrees. And in any case, it would not be a relevant measure of the similarity of the ratings of elements.

Then, combining the two ideas, the correlation coefficient has been chosen to represent the similarity. In this study, the Pearson correlation coefficient has been chosen. In our context the Pearson correlation coefficient has been named Affinity, so in advance whenever the Affinity term is mentioned it must be understood as the Pearson correlation coefficient.

Pearson correlation coefficients

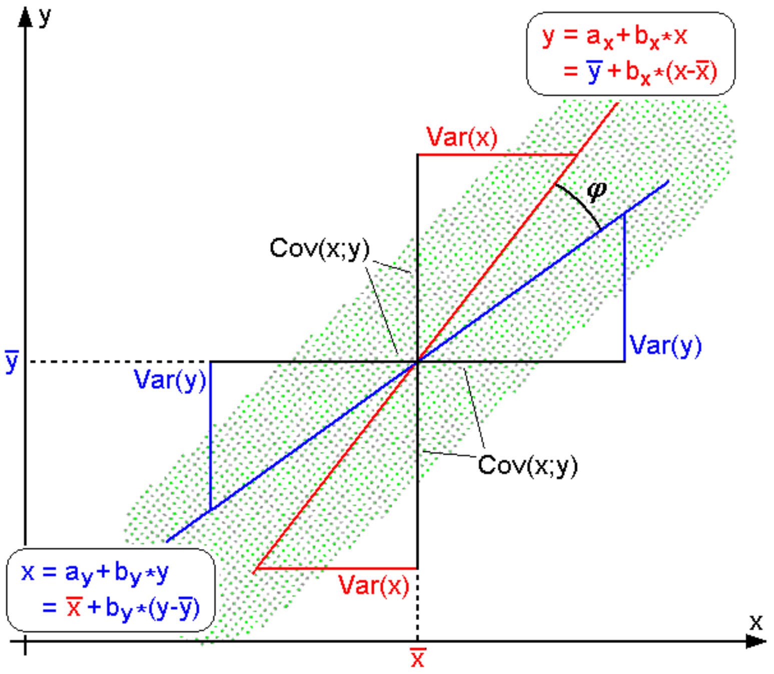
In order to measure the similitude between nodes, either users or movies, the correlation of the common ratings have been used. As correlation factor, the Pearson correlation factor has been used:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

In the case of Movies, to measure the correlation between two movies, all users that have rated both movies have been selected and the ratings have been used to create a distribution whose Pearson correlation has been extracted. The rates for movie 1 are the first set(X) on the distribution and the rates for movie 2 the second set(Y). The correlation factor will be thee weight between nodes in the projected network.

The Pearson coefficient is interpreted geometrically as the cosine among the two original distributions. For this to be true it is required that the two distributions are shifted to have a zero average. This process has not been applied because considering that we are handling distributions in the range 0 to 10 corresponding to the rates, the average of any two distributions will not be far from each other and the interpretation of the cosine distance will still apply. In any case, the magnitude will be a measure of similarity among nodes.

The coefficient correlation is a commutative function, so the obtained network is weighted undirected and without self-connections.



Projected movie network

With an initial number of 19777 movies, the fully connected projected network would have 400 million edges, making it difficult to handle while many of the edges would not be very representative as the number of common rates would be very low.

To address these two problems, a minimum number of rates have been required to calculate a correlation among the two nodes. This factor has been empirically to 31. No minimum correlation threshold has been set, obtaining then correlations in the range [-1, 1], where the typical value is from 0.4 to 0.6, not very discriminative but, enough to improve the rating prediction of the method, as it will be seen later.

The number of edges in the obtained network is 241936, around 2.4 edges on average by movie, although the distribution of nodes is quite unbalanced, as corresponds to a node that initially was following a Power Law distribution.

Projected user network

In the case of the user projection, having 274 users, a complete network could have been considered without obtaining a very high number of nodes, 75076, but the significance of the correlation would be biased towards high values when the number of common movies is low. To filter this issue, a number of 31 movies rated by the two users have also been required.

The number of edges in the obtained network is 37675, around half the total possible number of edges.

Small world effect

In order to evaluate the small world effect, it would have been required to have access to a larger number of users. If this had been the case, the correlation factor implemented for this work could have been applied and a different (extended) user graph would have been used to obtain recommendations.

It has not been possible to extract information from users others than soulmates from the website so it has not been possible to check how a different selection of users would affect the prediction of movie preferences and ratings.

The only correction introduced to measure user similarity is to require a minimum number of common ratings for two users have a connecting edge among them but this only affects user similarity as the users similar to my profile are already chosen.

Recommendation algorithm

With the aforementioned constructed networks, and based on [7] and [8] a method to conjugate the ratings of movies to predict the rating of movies not watched by the user has been defined.

The rating of a movie has been approximated by the weighted rating of the affine movies applying a trust coefficient that corresponds in our case to the Affinity.

$$r_{sm} = \frac{\sum_{i \in S} t_{si} r_{im}}{\sum_{i \in S} t_{si}}$$

The weight metric has been modified though in order to fine-tune the results, applying a modification, as described in [4]:

$$p_{ij} = m_i + \sum_{k=1}^{I} w_{ik}(r_{kj} - m_k)$$

p_{ij} is the prediction for movie i by user j

m_i is the average rating for the movie,

m_k is the average rating given to movie k

r_{kj} is the rate that user j gives to movie k

w_{ik} is a factor of similarity between movie i and movie k

The rationale behind this formula is that the characteristic feature of a user related to a movie is not its rating but the deviation to the average rating given by a large number of users ($r_{kj} - m_j$).

In order to compute the prediction of the rating all the affinities have been considered, no filtering on the magnitude of the correlation has been applied, even negative correlation have been included. This has required considering the absolute value of the affinity in the denominator of the formula.

The algorithm has been applied both considering similarities among users and movies.

In the case of movies, for each movie, the list of affine movies $\{m_j\}$ is extracted. For each of the movies its contribution to the rate to be predicted will be the difference of the user rate to the average rate of the movie multiplied by the affinity. The sum of the absolute value of the affinities is used to normalize the result. The result is subtracted to the average rate of the movie (μ_i).

$$r_i = \mu_i - \frac{\sum_j \ a_{ij} \left(\mu_j - r_j\right)}{\sum_j \left|a_{ij}\right|}$$

In the case of users the same function can be applied. In this case μ_j, represents the average rate given by the user to the movies rated by him and r_j the rate given by the user to the movie.

Rating reconstruction

To have an idea of the accuracy of the method proposed, the algorithm must be evaluated against a set of known results. In this study, the method will be used to reconstruct the ratings of the reference user. This will be done using two models. The first model has been generated considering the user ratings and the affine user ratings so, it is purely, a reconstruction of the rating as the result has been used to generate the model. In this case, both the user based similarity and the movie based similarity have been obtained. In this case, the system has been used also to measure how the reconstruction performs if instead of my own rates are reconstructed, the other users rates are.

The second model only considers the ratings given by the affine users, as extracted from the website, but the user ratings have not been introduced in the model. In this case, only the movie based similarity has been obtained as there is no correlation factor among my ratings and those of other users.

Baseline

In order to be able to evaluate the goodness of the algorithm, a baseline needs to be established. In this case, the baseline has been predicting the user rate as the average movie rate.

The result obtained can be seen in the following table for the 546 movies rated:

Mean Square Error	Mean Absolute Error	Linear Regression R factor	N
1.99141	1.149634	-0.79456	546

The values obtained for the error are close to 2 or the MSE, and 1.14 for the MAE. It is interesting also to notice that the R coefficient is highly negative, meaning that the user preferences are quite different from average.

Movie based Affinity

The affinity has been calculated based on the movie similarities. Initially, different thresholds for the affinity factor have been set. The result can be seen in the following table for two different cases, considering only positive or both positive and negative affinities.

MODE	THR	MSE	MAE	R	N	NMATCH
+	0.8	3.53	1.60	0.91	4	4
+	0.6	1.13	0.82	0.49	221	400
+	**0.4**	**0.82**	**0.68**	**0.72**	**464**	**9468**
+	0.2	0.93	0.75	0.73	471	57972
+	0	1.01	0.78	0.73	473	115260
-	0.8	3.53	1.60	0.91	4	4
-	0.6	1.13	0.82	0.49	221	400
-	**0.4**	**0.79**	**0.68**	**0.71**	**465**	**9556**

-	0.2	0.90	0.74	0.73	472	60088
-	0	0.97	0.77	0.73	473	136078

The first thing to notice is that the R factor has become strongly positive, what indicates that the prediction function is much closer to obtain accurate results.

A shocking aspect is that the results are much worst for the case where the affinity has been restricted to be over 0.8. Going into the detail, we observe that only 4 movies have another movie with a correlation factor over 0.8. It must be remembered, that this affinity is the affinity among all ratings of affine users. For 2 of the movies, namely ‘Dragon Ball’ and ‘Dragon Ball Z’, my rating is very different as I liked ‘Dragon Ball’ but disliked ‘Dragon Ball Z’ what is not the case for most users. In this case also the number of total movies used to calculate the MSE is very low, N=4 and each of them with only 1 match (NMATCH=1).

The results quickly improve when the threshold is lowered to 0.6. In this case, the MSE is 1.13 and MAE 0.82. In this case, a total of 221 movies have been rated, making the value much more representative. In this case, a significant improvement compared to the baseline is also observed.

If the threshold is reduced to 0.4 the results are further improved achieving MSE 0.82 and MAE 0.68, for 464 movies.

Interestingly, considering the negative correlation also helps to improve the accuracy by 0.03 for both 0.4 and 0.2 correlation threshold.

One limitation from this method is that not all movies have affinities, as they have not enough common ratings.

User based Affinity

To be able to provide a rate to every movie, the user based affinity algorithm is used. The errors obtained are shown in the next table:

THR	MSE	MAE	SR	N	NMATCH
0.4	**1.62**	**1.02**	**0.62**	**537**	**11317**
0.2	1.66	1.06	0.70	546	46634
0	1.66	1.06	0.70	546	48475

In this case, there are no affinity factors over 0.6 nor negative ones.

The results also improve the baseline but are significantly worse than the movie affinity based method.

MSE	MOVIE	USER
0.4	0.82	1.62
0.2	0.93	1.66

It provides in any case an alternative better than the baseline for those cases in which there is not enough movie affinities to obtain a prediction.

Movie based affinity without including own user rates

So far, the results obtained correspond to reconstruction errors as, in all cases, the rate that was being calculated was part of the data used to calculate the affinities. The result was then biased.

If the same model is constructed for movie affinities without using the rates of the user, the prediction results obtained are the ones shown in the following table:

MODE	THR	MSE	MAE	SR	N	NMATCH
+	0.8	3.90	1.68	0.91	4	4
+	0.6	1.85	1.04	0.60	215	394
+	**0.4**	**1.33**	**0.89**	**0.72**	**460**	**9260**
+	0.2	1.32	0.90	0.72	467	56474
+	0	1.32	0.91	0.72	468	111778
-	0.8	3.90	1.68	0.91	4	4
-	0.6	1.85	1.04	0.60	215	394
-	0.4	1.34	0.89	0.72	460	9354
-	0.2	1.32	0.90	0.72	467	58576
-	0	1.33	0.92	0.71	470	131854

Again the results are far better than the baseline established. The best results are obtained when the affinity threshold is set to 0.4, with MSE=1.33 and MAE=0.89.

In this case, the inclusion of negative correlation factors does not improve the accuracy.

Movie based affinity applied to all users.

If the model is applied to predict the rates of other users, the results obtained, considering an affinity threshold of 0.4 (the best obtained and that matches the best settings for the previous experiment) can be seen in the next table:

	AVG MSE	STD MSE	MIN MSE	AVG MAE	STD MAE	MIN MAE
AVERAGE	1.86	1.16	0.31	1.03	0.31	0.44
SOCIAL	2.05	1.21	0.57	1.08	0.29	0.63

In this case, the average prediction is better than the social prediction MSE=1.86 vs. 2.05, with a smaller standard deviation and minimum values, also the MAE is smaller.

This indicates that the choice of affine users is relevant, and that the set of chosen users even being the most affine users for the reference user, do not represent an appropriate set of affine users among them.

Results analysis

The obtained results confirm that the more information of the user included in the model, the better prediction is obtained. The correlation factor, affinity, is not effective as a criterion to improve the prediction.

A second odd effect is that in the case of movie based affinity, the consideration of this threshold reduces the chance of obtaining a prediction since a minimum number of coincident ratings have been required.

The affine user selection has proved to be very bounded to the user, as trying the same model for the rest of users in the model has obtained much worse results than for the reference user.

Movie recommendation

The final objective of the method is to be able to recommend new movies to the user. In order to select this, the recommended list of movies from each user is consolidated and added to the model. There are no ratings for any of the movies by the reference user.

The experiment is repeated for the model built including the user movie rates.

Recommendations considering user movie rates

The top 20 recommended movies, obtained with this method have been, according to the social rate, using the first mode have been:

Id	Title	Social Rate	Social ranking	Avg. Rate	Avg. Ranking	N Match
282386	Metropolis	9.56	1	8.2	25	1
667376	Witness for the Prosecution	9.11	2	8.7	1	6
874956	Game of Thrones (TV series)	9.09	3	8.5	4	14
645363	Donnie Darko	8.93	4	7.5	87	3
382807	**Les Misérables**	8.93	5	7.4	93	1
262430	Oldboy (Old Boy)	8.89	6	7.9	49	3
344683	The Kid	8.78	7	8.5	5	12
478505	Dersu Uzala	8.75	8	8.4	13	18
460960	How I Met Your Mother (TV Series)	8.75	9	7.8	58	13
536488	Sunset Boulevard	8.70	10	8.5	6	16
144674	Persepolis	8.62	11	7.8	59	6
791884	The Mission	8.59	12	7.6	73	14
168398	Seven Samurai	8.59	13	8.5	7	20
489970	Breaking Bad (TV Series)	8.58	14	8.5	8	20
802694	**Rear Window**	8.56	15	8.3	17	5
914839	The Man Who Shot Liberty Valance	8.55	16	8.4	14	8
831880	Boardwalk Empire (TV Series)	8.52	17	8.1	29	2
554741	**High Noon**	8.52	18	8.1	30	27
432676	Singin' in the Rain	8.51	19	8.1	31	23
381846	The Lives of Others	8.49	20	8.1	32	42

The ranking indicates the position when sorted according to the corresponding criteria. Although it cannot be verified, I can say that some of the movies recommended “Les Misérables”, “Rear Window” and “High noon” are movies that I had seen although I hadn’t rated them in the system.

If instead we sort according to the average rate:

Id	Title	Social Rate	Social ranking	Avg. Rate	Avg. Ranking	N Match
667376	Witness for the Prosecution	9.11	2	8.7	1	6
750301	City Lights	8.45	24	8.6	2	10
684718	To Be or Not to Be	8.02	66	8.6	3	10
874956	Game of Thrones (TV series)	9.09	3	8.5	4	14
344683	The Kid	8.78	7	8.5	5	12
536488	Sunset Boulevard	8.70	10	8.5	6	16
168398	Seven Samurai	8.59	13	8.5	7	20
489970	Breaking Bad (TV Series)	8.58	14	8.5	8	20
795770	The Apartment	8.32	34	8.5	9	8
640188	Band of Brothers (TV)	8.28	41	8.5	10	8
412004	City of God	8.21	48	8.5	11	14
448956	Paths of Glory	8.02	67	8.5	12	10
478505	Dersu Uzala	8.75	8	8.4	13	18
914839	The Man Who Shot Liberty Valance	8.55	16	8.4	14	8
402150	The Sopranos (TV Series)	8.35	31	8.4	15	24
875960	M	8.24	45	8.4	16	2
802694	Rear Window	8.56	15	8.3	17	5
357391	The Third Man	8.48	21	8.3	18	3
167667	Rebecca	8.48	22	8.3	19	11
701892	Apocalypse Now	8.29	38	8.3	20	13

A measure extracted from these rates is the average and standard deviation generated:

	AVG	STD
AVG RATE	7.45	0.67
SOCIAL RATE	7.68	0.76
AVG SOCIAL TOP 20	8.15	0.37
SOCIAL TOP 20	8.75	0.26
AVG TOP 20	8.46	0.11
SOCIAL AVG TOP 20	8.49	0.29

The average rate is higher for the social rate (calculated with the method) that the average rate of the movie. If only the TOP 20 ranked socially rated movies are considered, the difference in average rate grows. If instead the TOP 20 average rated is considered the average is almost the same. In both cases, the standard deviation gets reduced, as the range of rates is also reduced.

Conclusions and future work

The method implemented improves the prediction capabilities of the average rate. The best improvement obtained has been for a model where all the information available for the user was introduced, the method has been able to reconstruct the predictions with an average error of approximately of 0.68.

Similar methods studied, obtain accuracies of around 0.5(half star) but usually in a range of 0 to 6, while we are using a 0 to 10 scale.

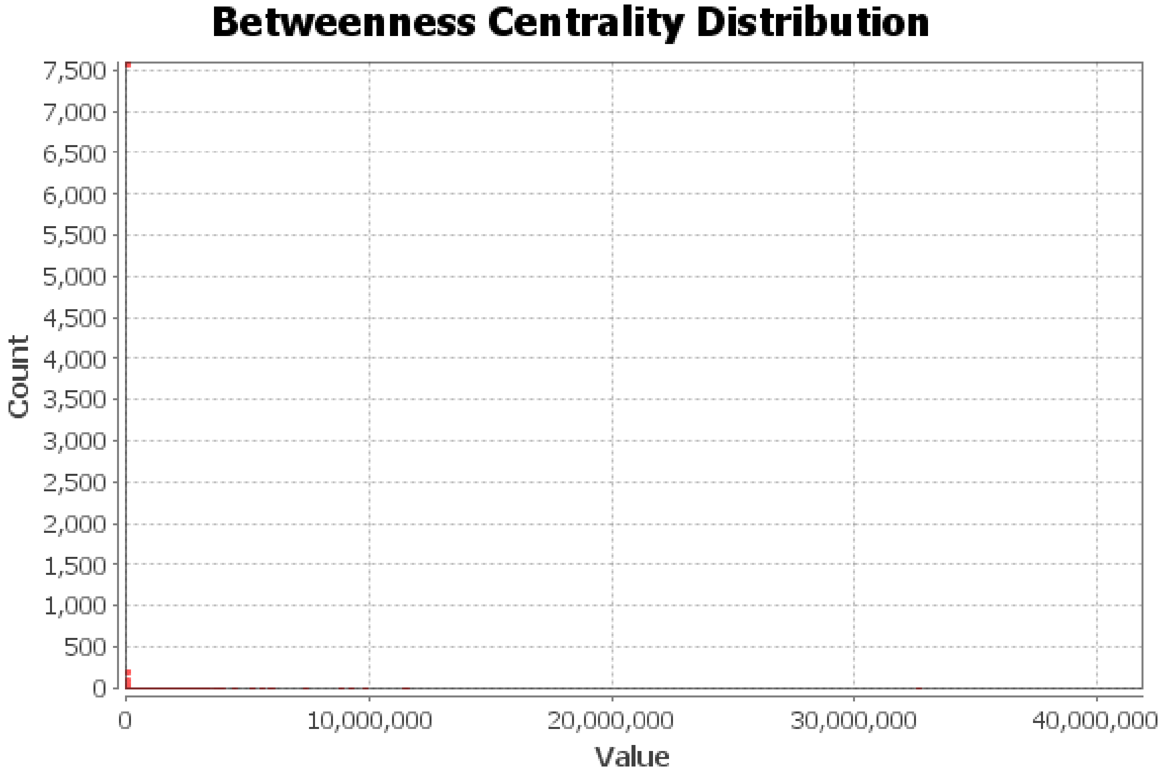
Given that the method benefits from including information specific to the user, additional parameters included in the movie and user information profile, like movie genre or user demographics (age, gender) could also help to improve the accuracy obtained. A combined correlation affinity could be extended from the current model.

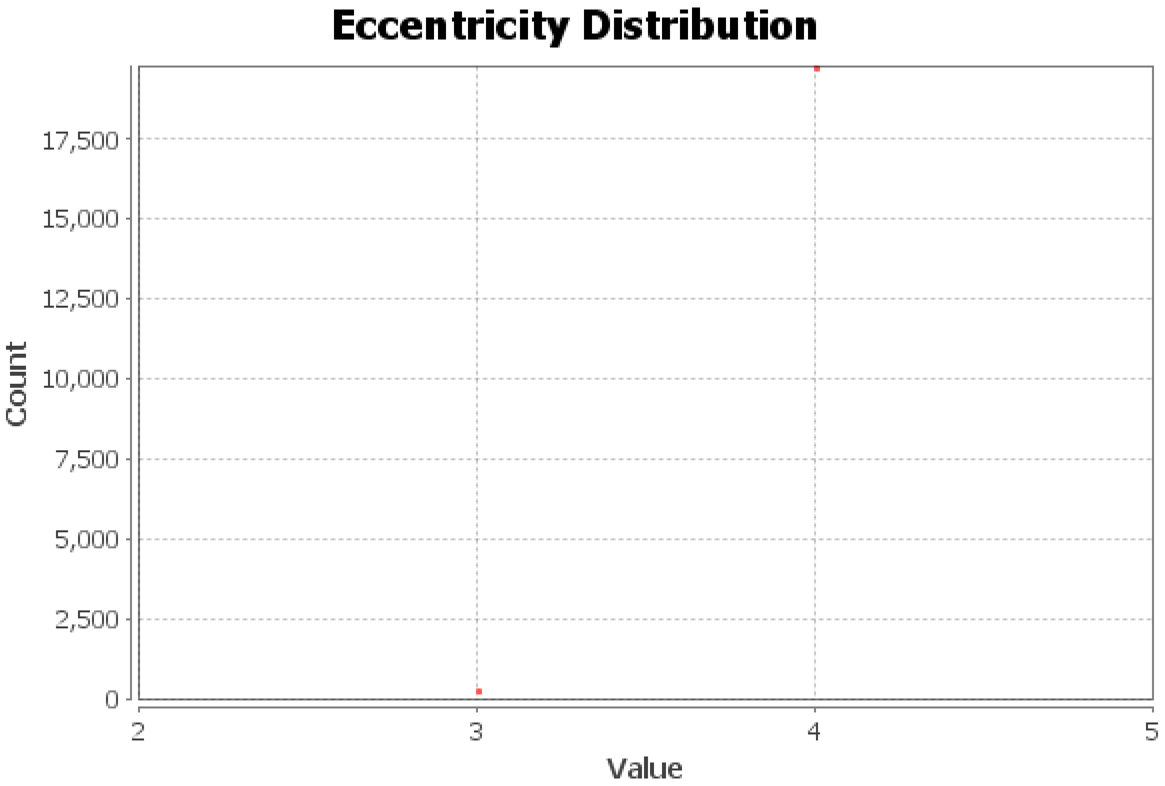
In any case, any change considered should consider a wider number of users as the users provided by Film Affinity are very limited, and the selection biased.

Appendix: Gephi user-movie rates graph analysis

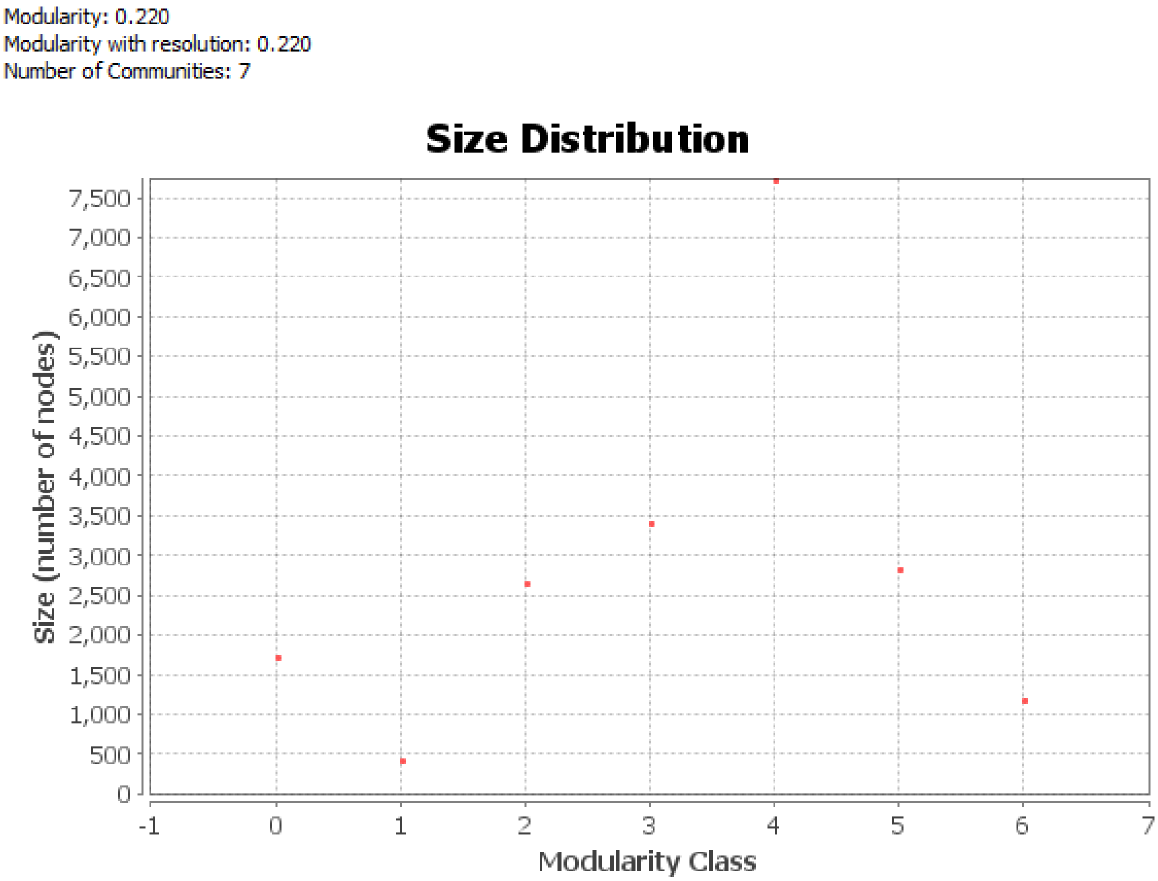
Some of the metrics that can be obtained automatically from Gephi for the network are shown.

Given the number of nodes, some graphs are not very illustrative. Where required in the analysis, an extraction of the data table has been shown.





The nodes with eccentricity 3 are users, while the nodes with eccentricity 4 are the movies.



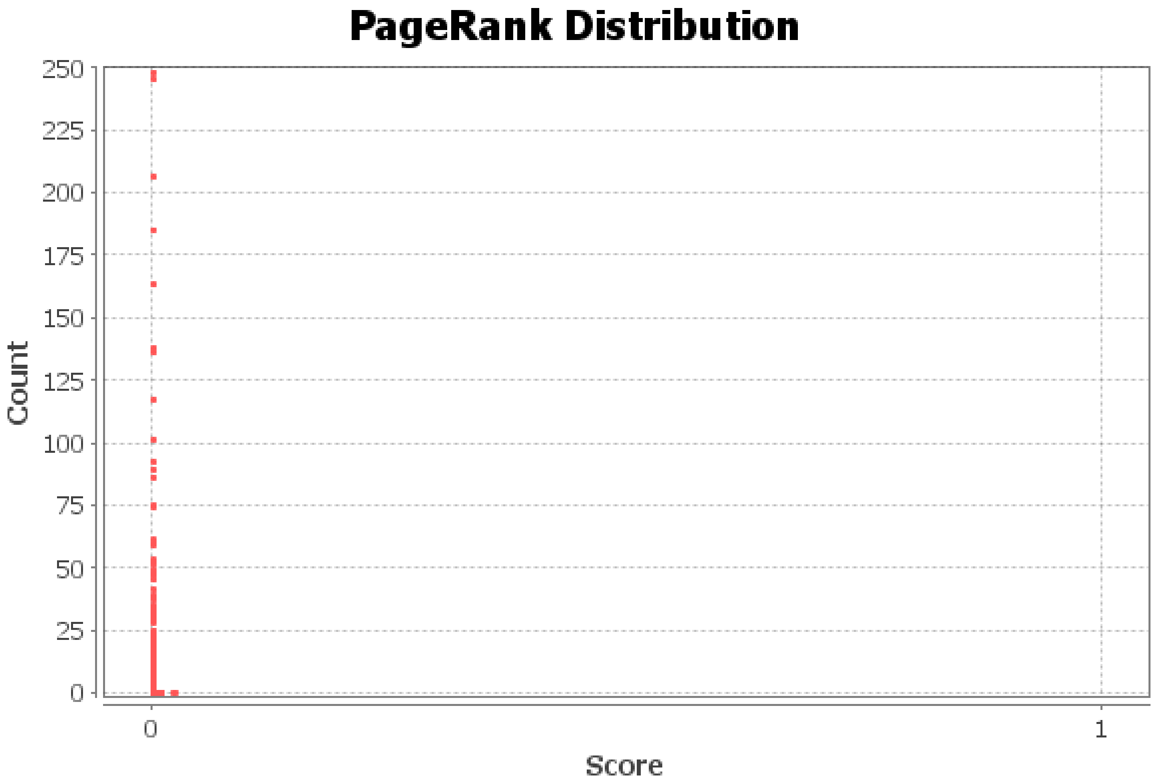
Modularity: 0.220
Modularity with resolution: 0.220
Number of Communities: 7

The modularity factor is low: 0.22, indicating that the clustering is not very compact.

Parameters:

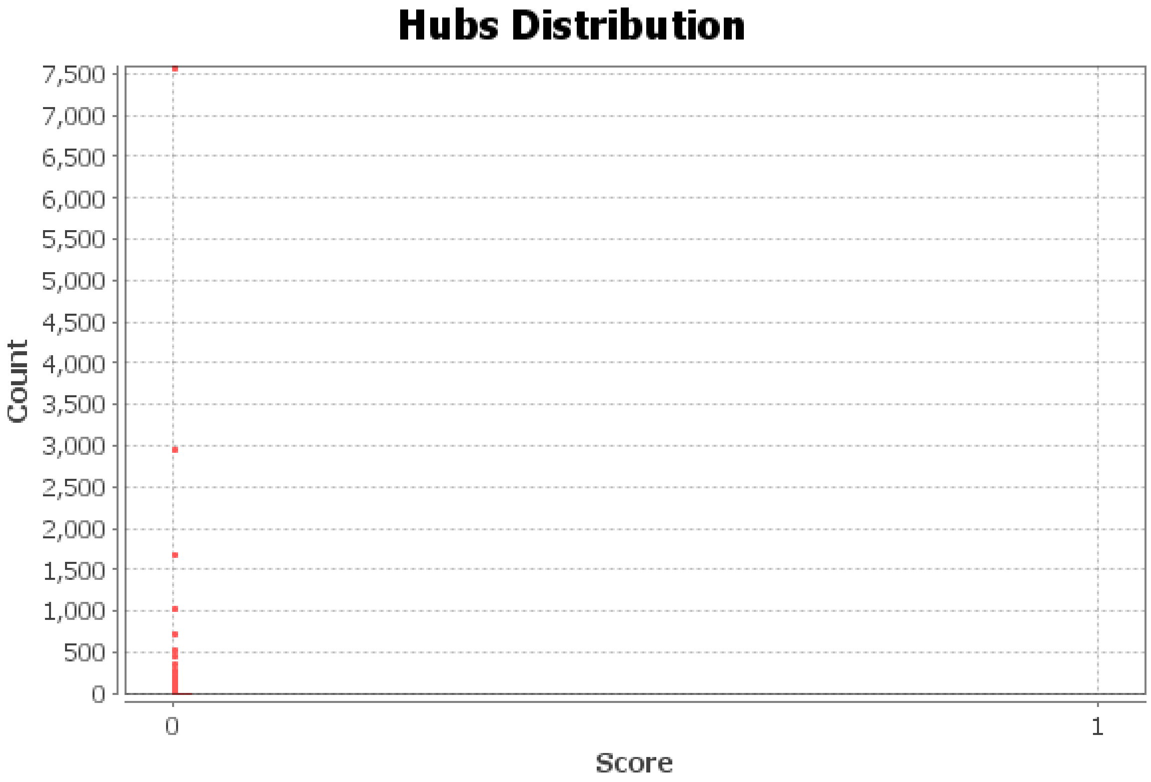
Epsilon = 0.001
Probability = 0.85

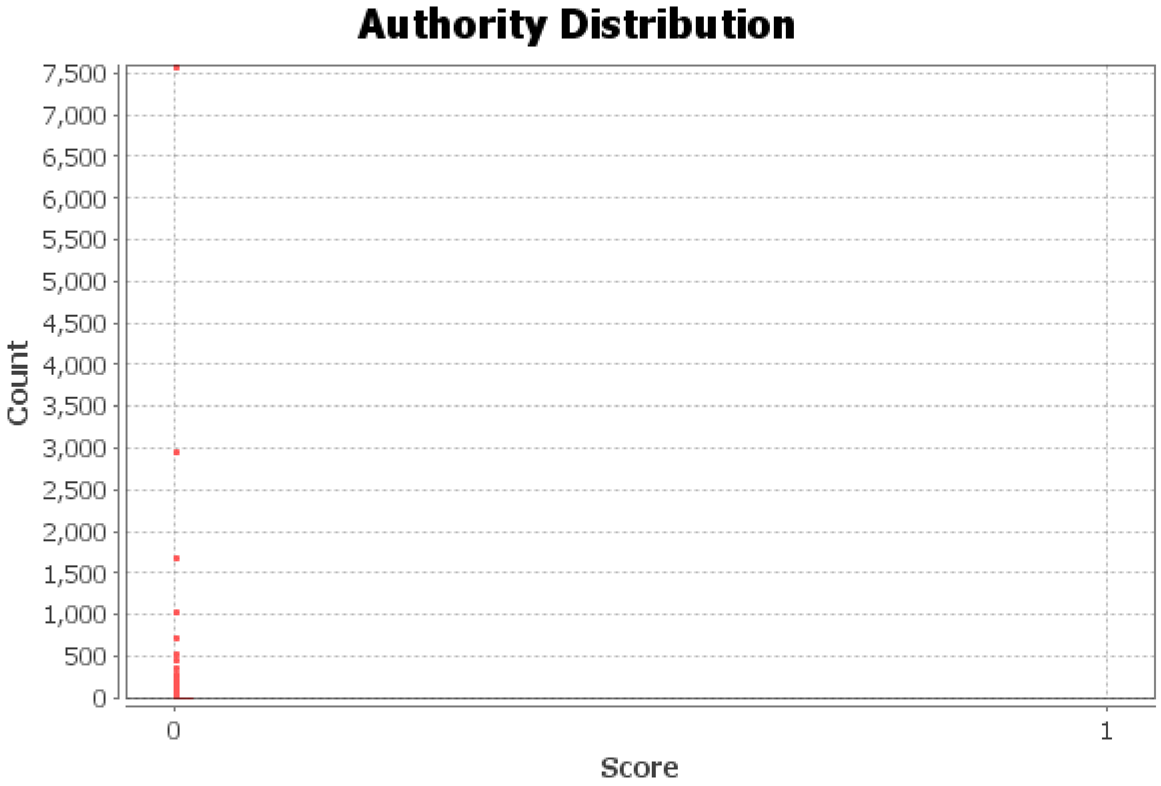
Results:



The agglomeration of the information does not allow to make an accurate analysis, but the function approximates the degree distribution of a Power Law, as has been seen from the analysis of the degree histogram.

Results:


Hubs Distribution
Count
7,500
7,000
6,500
6,000
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
0
1
Score


Authority Distribution
Count
7,500
7,000
6,500
6,000
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
0
1
Score

References

[1]MASSIMILIANO ZANIN, PEDRO CANO, JAVIER M. BULDÚ, OSCAR **CELMA Complex Networks in Recommendation Systems** , 2nd WSEAS Int. Conf on COMPUTER ENGINEERING and APPLICATIONS (CEA'08) Acapulco, Mexico, January 25-27, 2008
[2]Pedro Cano, Oscar Celma, and Markus Koppenberger, **The Topology of Music Recommendation Networks,** PACS numbers: 05.45.–a, 42.65.Sf, 42.55.Px, February 2, 2008.

[3]Walter, Battiston and Schweitzer**. A model of a Trust-based Recommendation System on a Social Network.** Revised manuscript for JAAMAS. September 2007**.**

[4]Ansari, Essegaier, Kohli**. Internet Recommendation Systems.**

[5]P Bonhard and M A Sasse**, 'Knowing me, knowing you' — using profiles and social networking to improve recommender systems.** BT Technology 84 Journal • Vol 24 No 3 • July 2006

[6]Jordi Palau, Miquel Montaner, Beatriz L´opez, and Josep Llu´ıs de la Rosa. **Collaboration Analysis in Recommender Systems using Social Networks.** Spanish MCYT Project DPI2001-2094-C03-01 and the Agentcities.NET Project ACNET.02.50.

[7]Jennifer Golbeck, James Hendler. **FilmTrust: Movie Recommendations using Trust in Web-based Social Networks**. Maryland Information and Network Dynamics Laboratory Semantic Web Agents Project,

[8]Jennifer Golbeck. **Generating Predictive Movie Recommendations from Trust in Social Networks.** Maryland Information and Network Dynamics Laboratory Semantic Web Agents Project.

[9]Quentin Jones and Sukeshini A. Grandhi, **P3 Systems:Putting the Place Back into Social Networks.** SEPTEMBER • OCTOBER 2005 Published by the IEEE Computer Society. 1089-7801/05

[10]Kautz, Selman and Shah. **Combining Social Networks and Collaborative Filtering**. COMMUNICATIONS OF THE ACM March 1997/Vol 40. No. 3

[11]Tao Zhou, Jie Ren,Matúš Medo and Yi-Cheng Zhang1. **Bipartite network projection and personal recommendation,** Physical Review E 76(4): 046115, 2007

[12]Jacob Goldenberg, Barak Libai, Eitan Muller. **Talk of the network, a complex systems look at the underlying process of word-of-mouth.** Marketing Letters 12:3, 211-223, 2001