# Some Approaches to the Analysis and Visualization of the Internet Movie Database

Vladimir Batagelj and Andrej Mrvar
University of Ljubljana, Slovenia

Adel Ahmed, Xiaoyan Fu, Seok-Hee Hong and Damian Merrick
National ICT Australia, Sydney, Australia

September 8, 2005

The source of the original data is the Internet Movie Database.

We transformed the contest data into a **Pajek** temporal network with some additional vectors and partitions describing the properties of vertices.

```
imdb.net    - imdb network in Pajek format
imdbL.net   - imdb network with long names
imdb.clu    - type partition of vertices
imdb.nam    - long names
imdb.vec    - year
large.net   - largest weak component with long names
large.vec   - years for large
largeT.clu  - type partition for large
largeB.clu  - bipartition for large
```

The file imdb.clu contains the following classes:

```
0 Actor             11 Crime
1 Drama             12 Sci-Fi
2 Short             13 Horror
3 Documentary       14 War
4 Comedy            15 Fantasy
5 Western           16 Romance
6 Family            17 Adventure
7 Mystery           18 Animation
8 Thriller          19 Action
9 Adult             20 Musical
10 Music            21 Film-Noir
99 Unknown
```

The **Pajek** data files are available at **Pajek** 's data sets page.

# Basic characteristics of IMDB

The IMDB network is bipartite (2-mode) and has $1324748 = 428440 + 896308$ vertices and 3792390 arcs.

9927 of the arcs in the network are multiple (parallel) arcs. Here is their distribution.

| multiplicity | frequency |
|---|---|
| 1 | 3775126 |
| 2 | 6178 |
| 3 | 588 |
| 4 | 267 |
| 5 | 128 |
| 6 | 66 |
| 7 | 45 |
| 8 | 18 |
| 9 | 23 |
| 10 | 6 |
| 11 | 5 |
| 12 | 3 |
| 13 | 2 |
| 15 | 1 |
| 16 | 1 |
| 17 | 2 |
| 22 | 1 |
| 32 | 1 |
| 35 | 1 |
| 43 | 1 |

The nature of the appearance of multiple arcs can be seen from the Figure 1 where all arcs with multiplicity at least 8 are displayed.

**In the analyses that follow, we decided to treat multiple arcs as single.**

The IMDB network consists of 132714 weak components. Here is the distribution of their sizes.

| Size | Freq | Size | Freq |
|---|---|---|---|
| 1 | 124829 | 21 | 9 |
| 2 | 3557 | 22 | 3 |
| 3 | 1526 | 23 | 6 |
| 4 | 922 | 24 | 4 |
| 5 | 615 | 25 | 5 |
| 6 | 424 | 26 | 2 |
| 7 | 219 | 27 | 1 |
| 8 | 139 | 28 | 1 |
| 9 | 107 | 29 | 1 |
| 10 | 80 | 31 | 4 |
| 11 | 67 | 32 | 1 |
| 12 | 43 | 33 | 1 |
| 13 | 28 | 35 | 1 |
| 14 | 31 | 37 | 2 |
| 15 | 15 | 40 | 1 |
| 16 | 19 | 42 | 1 |
| 17 | 10 | 45 | 2 |
| 18 | 16 | 50 | 1 |
| 19 | 12 | 58 | 1 |
| 20 | 6 | 73 | 1 |
| 21 | 9 | 1169725 | 1 |

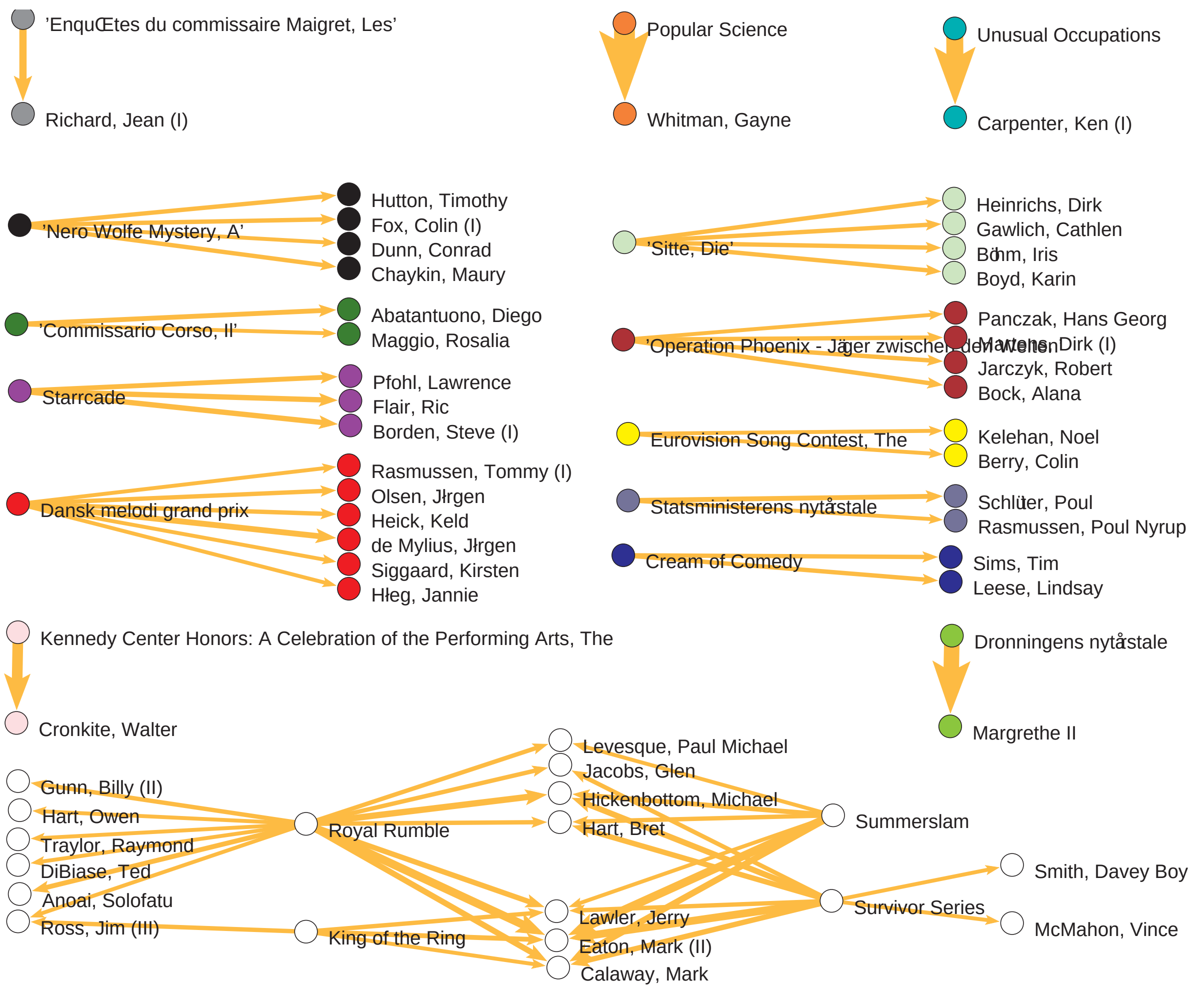



Figure 1: *Arcs with multiplicity at least 8*

# Identifying interesting parts of bipartite networks

There are few direct specialized methods for analyzing bipartite (2-mode) networks, especially large ones. Also, because of the size of the IMDB network, the standard reduction of the entire network to one or the other derived 1-mode network was not an option. The only special method available in **`Pajek`** was the adapted version of *hubs and authorities*, which did not produce very interesting results. We started to think about some new methods. Last August we developed and implemented in **`Pajek`** two new methods for analysis of bipartite networks:

- bipartite version of cores – $(p, q)$-cores
- 4-rings weights on lines

For details see Dagstuhl seminar 05361 / Batagelj.

## $(p, q)$-cores

The subset of vertices $C \subseteq V$ is a *$(p, q)$-core* in a bipartite (2-mode) network $N = (V_1, V_2; L)$, $V = V_1 \cup V_2$ iff

**a**. in the induced subnetwork $K = (C_1, C_2; L(C))$, $C_1 = C \cap V_1$, $C_2 = C \cap V_2$ it holds $\forall v \in C_1 : \deg_K(v) \geq p$ and $\forall v \in C_2 : \deg_K(v) \geq q$ ;

**b**. $C$ is the maximal subset of $V$ satisfying condition **a**.

The basic properties of bipartite cores are:

- $C(0, 0) = V$
- $K(p, q)$ is not always connected
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_1, q_1) \subseteq C(p_2, q_2)$

There exists a very efficient $O(m)$ algorithm to determine $(p, q)$-cores.

Since there are many $(p, q)$-cores, we must answer the question of how to select the interesting ones among them. To help the user in these decisions, we implemented in **`Pajek`** a *Table of cores' characteristics* $n_1 = |C_1(p, q)|$, $n_2 = |C_2(p, q)|$ and $k$ – number of components in $K(p, q)$. We look for $(p, q)$-cores where

- $n_1 + n_2 \leq$ selected threshold
- big jumps from $C(p - 1, q)$ and $C(p, q - 1)$ to $C(p, q)$.

We selected (247,2)-core, (27,22)-core and (2,516)-core. From the labels we can see that the corresponding topics are wrestling and pornography.

Table 1: $(p, q : n_1, n_2)$ for IMDB

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1590: | 1590 | 1 | 22 | 24: | 1854 | 1153 | 43 | 14: | 29 | 83 |
| 2 | 516: | 788 | 3 | 23 | 23: | 47 | 56 | 44 | 14: | 29 | 83 |
| 3 | 212: | 1705 | 18 | 24 | 23: | 34 | 39 | 45 | 13: | 30 | 95 |
| 4 | 151: | 4330 | 154 | 25 | 22: | 42 | 53 | 46 | 13: | 29 | 94 |
| 5 | 131: | 4282 | 209 | 26 | 22: | 31 | 38 | 47 | 12: | 29 | 101 |
| 6 | 115: | 3635 | 223 | 27 | 22: | 31 | 38 | 48 | 12: | 28 | 100 |
| 7 | 101: | 3224 | 244 | 28 | 20: | 36 | 53 | 49 | 12: | 26 | 95 |
| 8 | 88: | 2860 | 263 | 29 | 20: | 35 | 52 | 50 | 11: | 27 | 111 |
| 9 | 77: | 3467 | 393 | 30 | 19: | 35 | 59 | 51 | 11: | 26 | 110 |
| 10 | 69: | 3150 | 428 | 31 | 19: | 35 | 59 | 52 | 11: | 16 | 79 |
| 11 | 63: | 2442 | 382 | 32 | 19: | 34 | 57 | 53 | 10: | 35 | 162 |
| 12 | 56: | 2479 | 454 | 33 | 18: | 34 | 62 | 54 | 10: | 35 | 162 |
| 13 | 50: | 3330 | 716 | 34 | 18: | 34 | 62 | 55 | 10: | 34 | 162 |
| 14 | 46: | 2460 | 596 | 35 | 18: | 33 | 61 | 56 | 10: | 34 | 162 |
| 15 | 42: | 2663 | 739 | 36 | 17: | 33 | 65 | 57 | 9: | 35 | 187 |
| 16 | 39: | 2173 | 678 | 37 | 16: | 33 | 75 | 58 | 9: | 33 | 180 |
| 17 | 35: | 2791 | 995 | 38 | 16: | 30 | 73 | 59 | 9: | 33 | 180 |
| 18 | 32: | 2684 | 1080 | 39 | 16: | 29 | 70 | 60 | 9: | 32 | 178 |
| 19 | 30: | 2395 | 1063 | 40 | 15: | 29 | 77 | 61 | 9: | 31 | 177 |
| 20 | 28: | 2216 | 1087 | 41 | 15: | 28 | 76 | 62 | 9: | 31 | 177 |
| 21 | 26: | 1988 | 1087 | 42 | 15: | 28 | 76 | 63 | 8: | 31 | 202 |

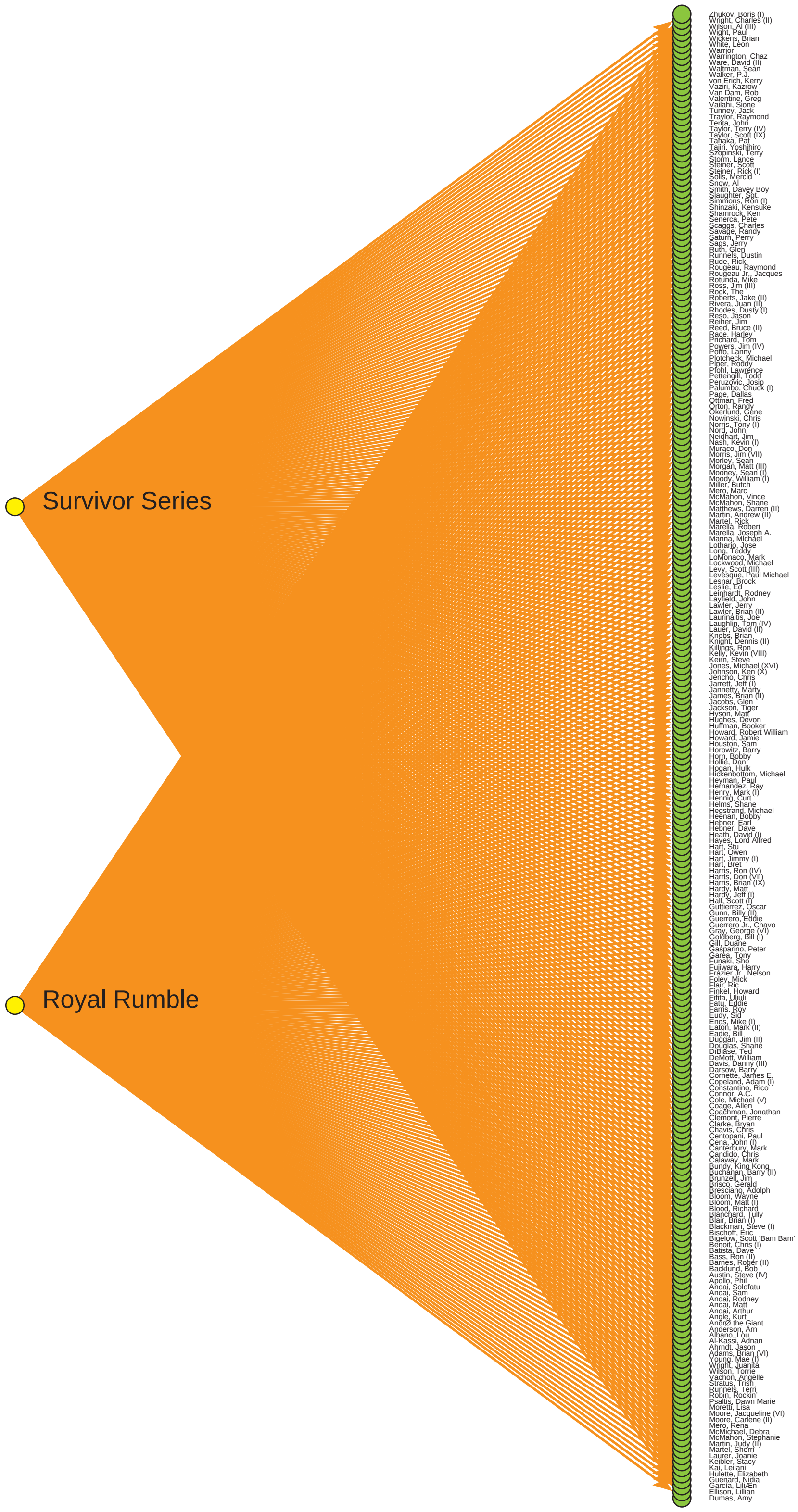



Figure 2: *(247,2)-core*


Figure 3: *(27,22)-core*

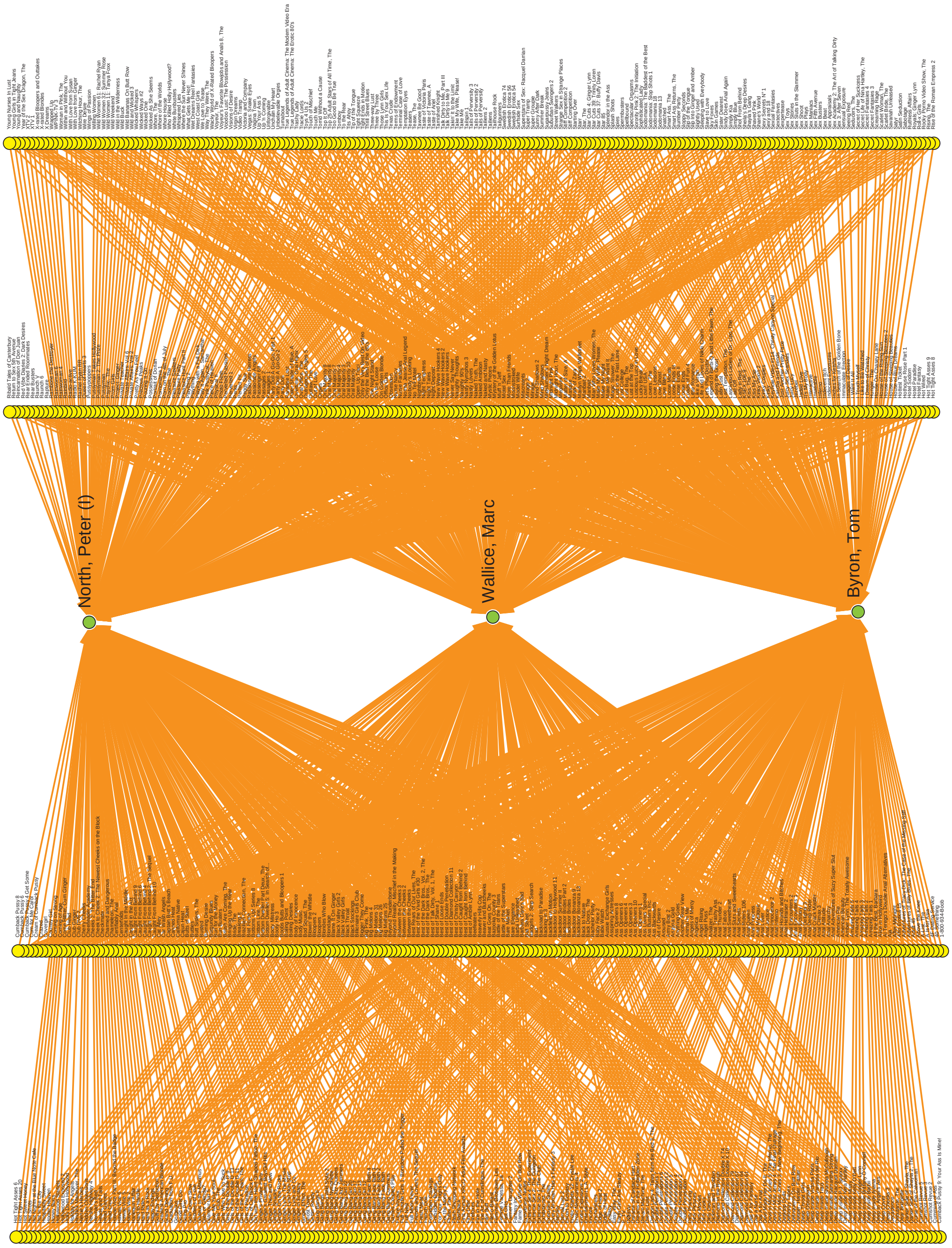
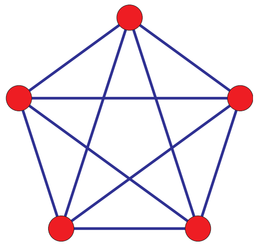



Figure 4: *(2,516)-Hard core*

## 4-rings

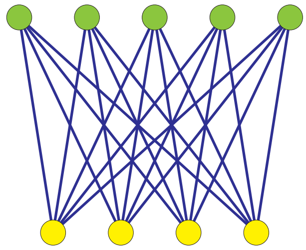
A *k-ring* is a simple closed chain of length $k$. Using $k$-rings we can define a weight of edges as
$w_k(e) = \#$ of different $k$-rings containing the edge $e \in E$



Since for a complete graph $K_r$, $r \geq k \geq 3$ we have $w_k(K_r) = (r-2)!/(r-k)!$ the edges belonging to cliques have large weights. Therefore these weights can be used to identify the dense parts of a network.

For example: all $r$-cliques of a network belong to $r-2$-edge cut for the weight $w_3$.

The 3-rings weights were implemented in **Pajek** in May 2002. However, there are no 3-rings in the IMDB network. The densest substructures are complete bipartite subgraphs $K_{p,q}$. They contain many 4-rings.



$$w_4(K_{p,q}) = (p-1)(q-1)$$

So we decided to implement 4-rings weights in **Pajek** .

To identify interesting substructures we applied the simple islands procedure for the weight $w_4$. It takes around 3 minutes to compute $w_4$ weights on a 1400 MHz, 1GB RAM computer, and 13 seconds to determine the islands.

We obtained 12465 simple line islands on 56086 vertices. Here is their size distribution.

There are 94 of size at least 30; and only 10 over 100. Again the largest island corresponds to wrestling. Each island represents a special topic. We visualized only some of them.

| Island | Size | Representative |
|---|---|---|
| 1 | 673 | Andre the Giant: Larger Than Life |
| 2 | 332 | 13. jul |
| 3 | 332 | Aa bakudan |
| 4 | 301 | Aa Chithrasalabham Parannotte |
| 5 | 269 | Adult 45 |
| 6 | 163 | .hack//Akusei heni vol. 2 |
| 7 | 144 | Aladdin |
| 8 | 135 | Gondoliers, The |
| 9 | 122 | Bag om Robinson ekspeditionen |
| 10 | 106 | 1992 Winter Olympics Figure Skating |

Table 2: $(p, q : n_1, n_2)$ for IMDB

| Size | Freq | Size | Freq | Size | Freq | Size | Freq |
|---|---|---|---|---|---|---|---|
| 2 | 5512 | 20 | 19 | 38 | 4 | 59 | 2 |
| 3 | 1978 | 21 | 18 | 39 | 3 | 61 | 1 |
| 4 | 1639 | 22 | 15 | 40 | 2 | 64 | 1 |
| 5 | 968 | 23 | 9 | 42 | 2 | 67 | 1 |
| 6 | 666 | 24 | 13 | 43 | 3 | 70 | 1 |
| 7 | 394 | 25 | 12 | 45 | 3 | 73 | 1 |
| 8 | 257 | 26 | 6 | 46 | 4 | 76 | 1 |
| 9 | 209 | 27 | 6 | 47 | 5 | 82 | 1 |
| 10 | 148 | 28 | 5 | 48 | 1 | 86 | 1 |
| 11 | 118 | 29 | 6 | 49 | 2 | 106 | 1 |
| 12 | 87 | 30 | 3 | 50 | 2 | 122 | 1 |
| 13 | 55 | 31 | 6 | 51 | 1 | 135 | 1 |
| 14 | 62 | 32 | 5 | 52 | 2 | 144 | 1 |
| 15 | 46 | 33 | 3 | 53 | 1 | 163 | 1 |
| 16 | 39 | 34 | 1 | 54 | 2 | 269 | 1 |
| 17 | 27 | 35 | 5 | 55 | 1 | 301 | 1 |
| 18 | 28 | 36 | 4 | 57 | 1 | 332 | 2 |
| 19 | 29 | 37 | 7 | 58 | 1 | 673 | 1 |

| | | |
|---|---|---|
| 11 | 86 | Accouplements pour voyeurs |
| 12 | 82 | Affren i Mlleby |
| 13 | 76 | Emmanuelle Forever |
| 14 | 73 | Directing Rye |
| 15 | 70 | 002 agenti segretissimi |
| 16 | 67 | Adventures of Red Ryder |
| 17 | 64 | Abuse me... 1: Feuchte Pppchen |
| 18 | 61 | Real World Reunion 2000, The |
| 19 | 59 | Abid el gassad |
| 20 | 59 | Jiyu gakkou |
| 21 | 58 | IDandT Presents the Darkraver |
| 22 | 57 | All Around Cure, An |
| 23 | 55 | AandE Biography: John Waters |
| 24 | 54 | Avonturen van een zigeunerjongen |
| 25 | 54 | All Aboard |
| 26 | 53 | Adventures of Mark Twain, The |
| 27 | 52 | Binge and Purge |
| 28 | 52 | Aladdin’s Lantern |
| 29 | 51 | Survivor - Season One: The Greatest and Most Outrageous Moments |
| 30 | 50 | Polizeiruf 110 - Angst um Tessa Blow |
| 31 | 50 | Abouna |
| 32 | 49 | Kid senshi Gundam: Meguriai sora |
| 33 | 49 | Buster Be Good |
| 34 | 48 | Auf ins blaukarierte Himmelbett |
| 35 | 47 | Accident, L’ |
| 36 | 47 | Adventures of Elmo in Grouchland, The |
| 37 | 47 | Eurovision Song Contest, The |
| 38 | 47 | Beaches |
| 39 | 47 | Bubblegum Crisis Tokyo 2040: Shadow War |
| 40 | 46 | Bingville Fire Department, The |
| 41 | 46 | Advoktka Vera |
| 42 | 46 | Angel of Destruction |
| 43 | 46 | Cry in the Dark, A |
| 44 | 45 | Lawrence Welk: Milestones and Memories - A Musical Family Reunion |
| 45 | 45 | Millennium Madness: Gangbangers of America |
| 46 | 45 | Zombie Planet |
| 47 | 43 | Polizeiruf 110 - Abschiedslied fr Linda |
| 48 | 43 | Ali Baba bujang lapok |

| | | |
|---|---|---|
| 49 | 43 | Entfhrung aus der Lindenstrae |
| 50 | 42 | Stained Memories |
| 51 | 42 | Helden von Bern, Die |
| 52 | 40 | Berlin Snuff |
| 53 | 40 | Amerikaansche meisjes |
| 54 | 39 | Atunci i-am condamnat pe toti la moarte |
| 55 | 39 | Tatort - ... und die Musi spielt dazu |
| 56 | 39 | Dalziel and Pascoe: A Clubbable Woman |
| 57 | 38 | Beszl knts, A |
| 58 | 38 | Ahasin Polawatha |
| 59 | 38 | Undressed: The Casting Couch |
| 60 | 38 | Aladim e a Lmpada Maravilhosa |
| 61 | 37 | Doppelter Einsatz - Auf Leben und Tod |
| 62 | 37 | Miss Belgi 1994 |
| 63 | 37 | 'Bar' |
| 64 | 37 | Una y media |
| 65 | 37 | 'Club de Los Tigritos, El' |
| 66 | 37 | Easter Carol, An |
| 67 | 37 | Carmen, a cigana |
| 68 | 36 | Abuelo, la condesa y Escarlata la traviesa, El |
| 69 | 36 | Be My Valentine, Charlie Brown |
| 70 | 36 | Hei kliffaa hei! |
| 71 | 36 | Carry On Abroad |
| 72 | 35 | 'Brug, De' |
| 73 | 35 | Escape Through Time |
| 74 | 35 | Et la lumire fut |
| 75 | 35 | Paper-Thin Immortals |
| 76 | 35 | Best of Big Brother, The |
| 77 | 34 | A los cirujanos se les va la mano |
| 78 | 33 | 'Shortland Street' |
| 79 | 33 | Jri Rumm |
| 80 | 33 | Boys to Men |
| 81 | 32 | Amor de Perdio |
| 82 | 32 | Circo de las Montini, El |
| 83 | 32 | Newlyweds Build, The |
| 84 | 32 | Alice at the Carnival |
| 85 | 32 | Bulle von Tlz - Bauernhochzeit, Der |
| 86 | 31 | 'Fugitivos Reality Mission' |
| 87 | 31 | Dark Area, The |
| 88 | 31 | Boh fett |
| 89 | 31 | Secret Spot, The |
| 90 | 31 | Heftig og begeistret |
| 91 | 31 | AandE Biography: Stooges - The Men Behind the Mayhem |
| 92 | 30 | Aliki dictator, I |
| 93 | 30 | Cabaret! |
| 94 | 30 | Andel's Story |
| 95 | 29 | Abnormal Man |

---

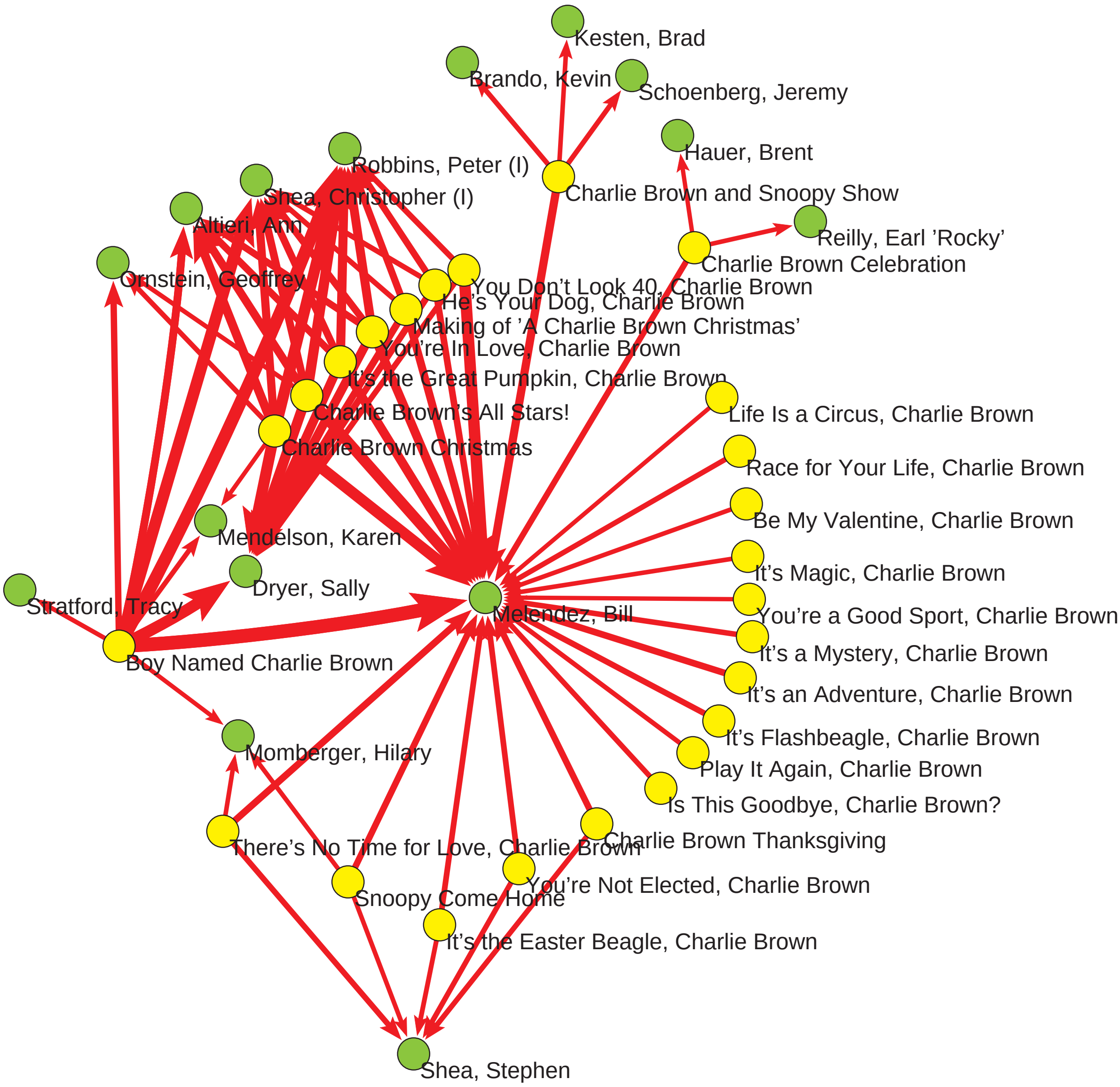



Figure 5: *Charlie Brown*

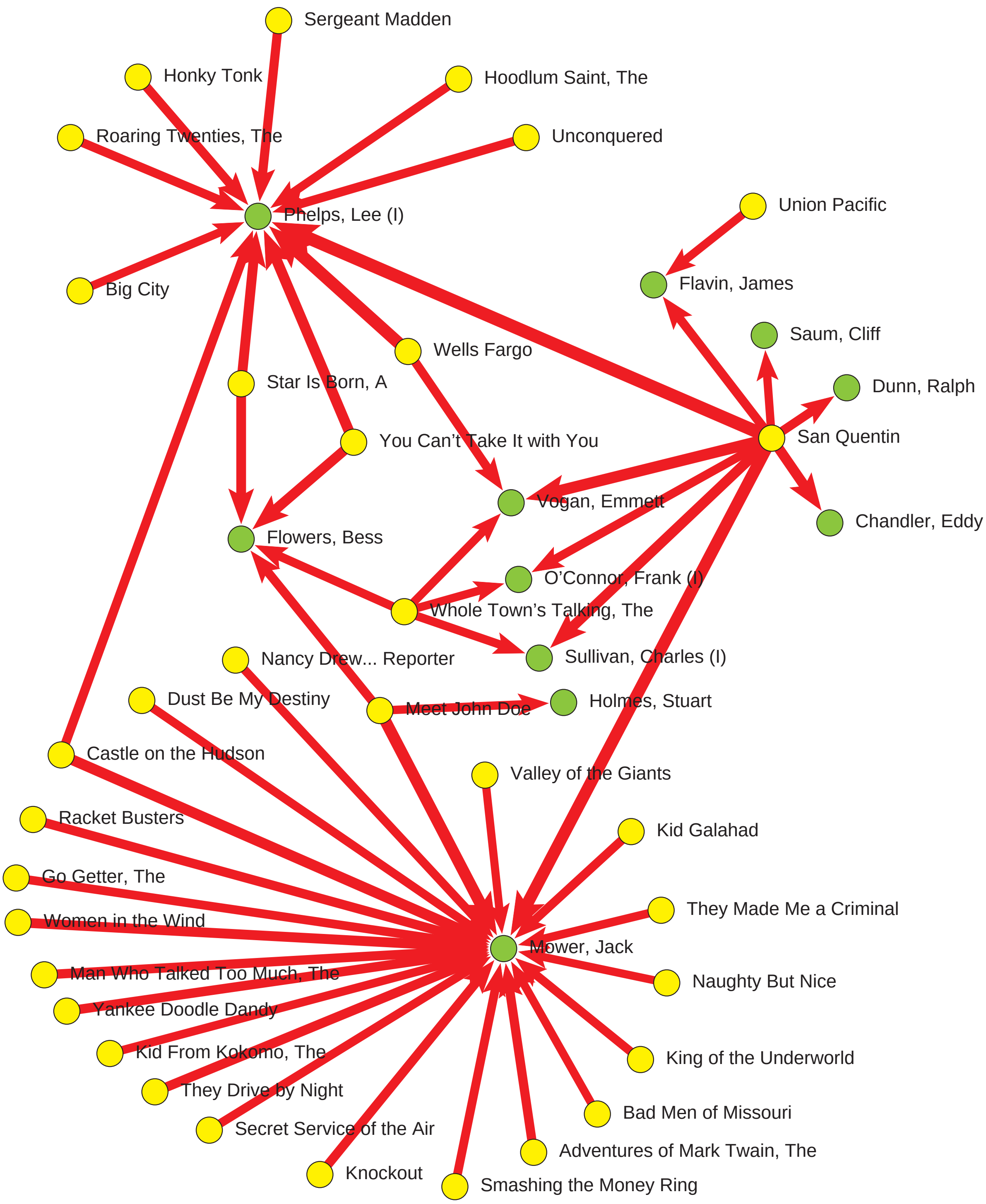



Figure 6: *Mower, Jack and Phelps, Lee*

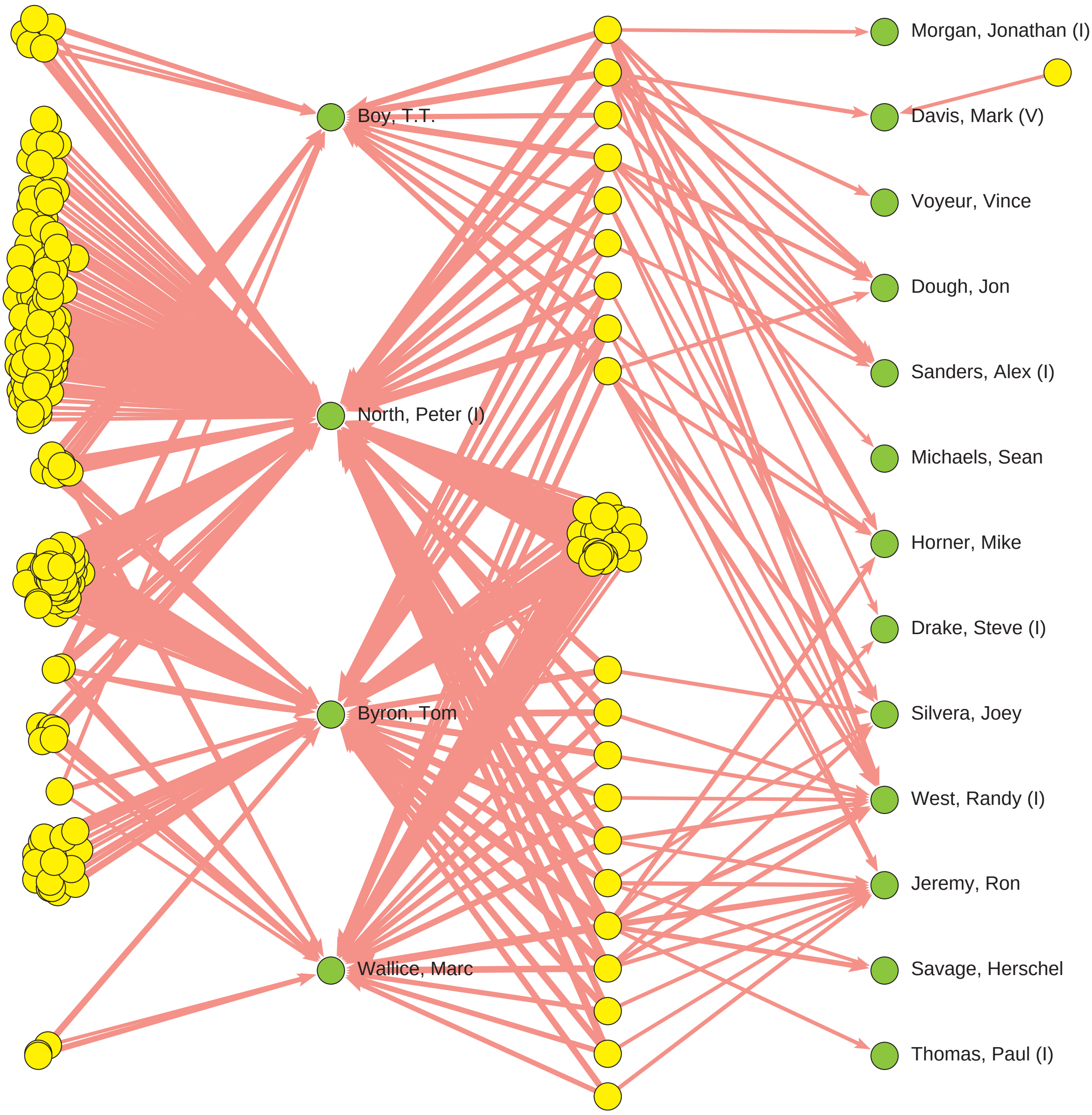



Figure 7: *Adult*

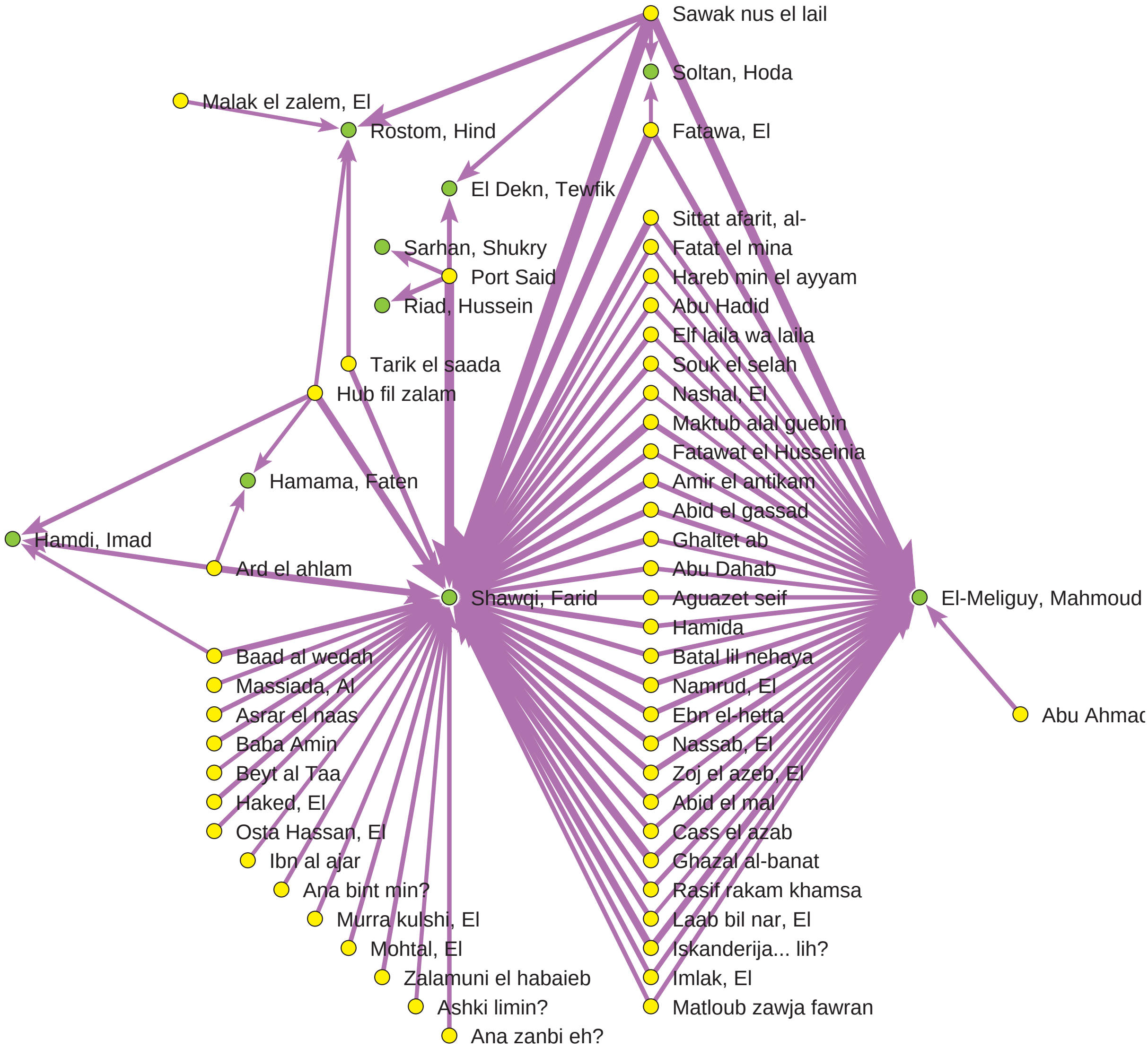



Figure 8: *Shawqi, Farid and El-Meliguy, Mahmoud*

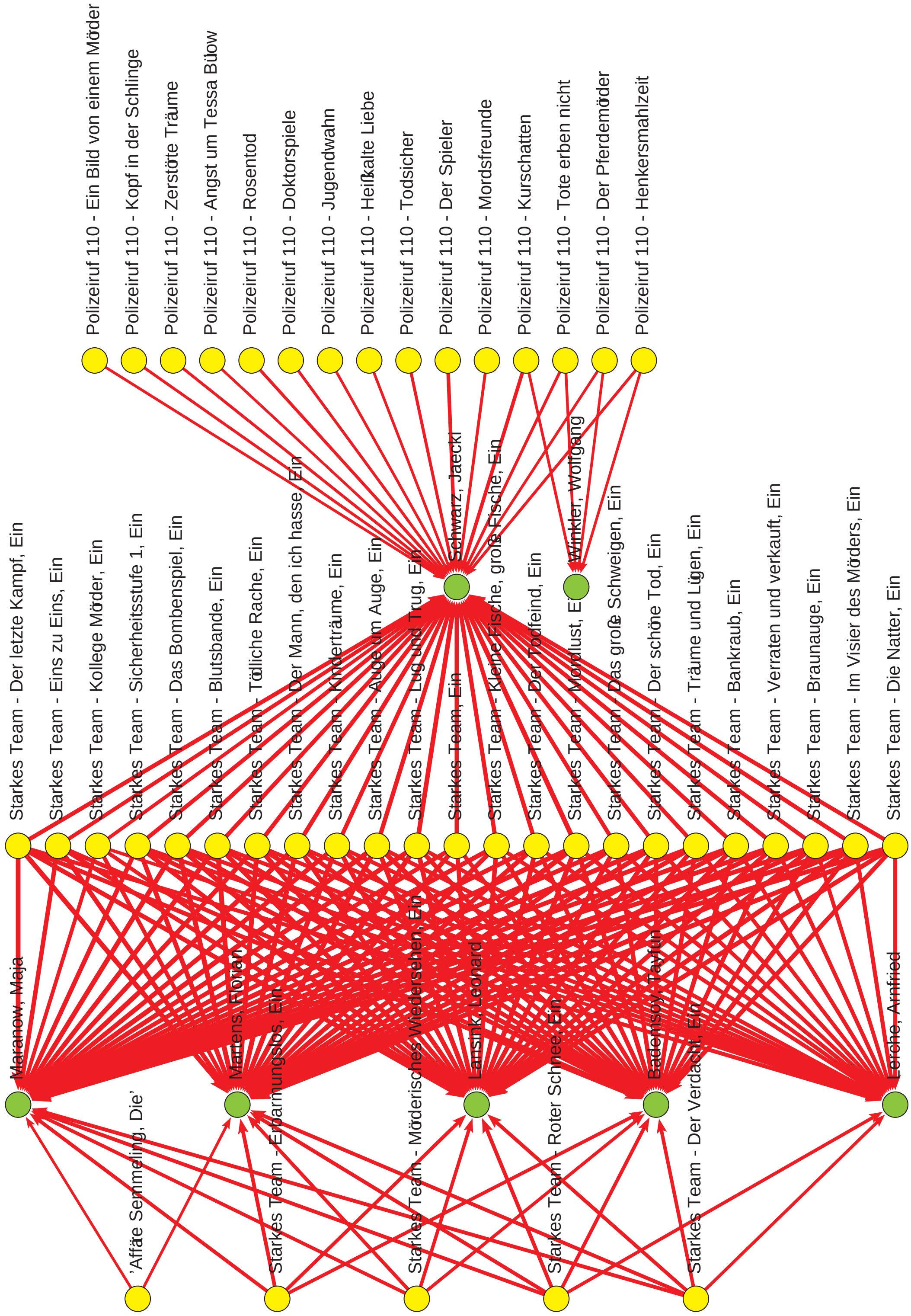



Figure 9: *Polizeiruf 110 and Starkes Team*

## Time slices

By extracting a time slice from the complete network, we can identify the main groups in selected time periods. To illustrate this, we extracted the time slice 1935-1950.

There are 223 simple islands for $w_4$ on 1774 vertices.

| Island | Size | Representative |
|---|---|---|
| 1 | 139 | ABC Mark Curry & Delta Burke Back Lot Special |
| 2 | 85 | Bag klosterets mure |
| 3 | 73 | Kaikki peliss |
| 4 | 73 | A Yong |
| 5 | 64 | Bartom, Bdy Gbor |
| 6 | 63 | Doa Macabra |
| 7 | 49 | Ako-Jo danzetsu |
| 8 | 38 | Gubernator |
| 9 | 35 | Dancing on the Face of the Moon |
| 10 | 35 | Mdgmurebi |
| 11 | 25 | Dandy Dan - He’s a Detective |
| 12 | 25 | Barrister Parvatishan |
| 13 | 24 | Abas Largas, Os |
| 14 | 23 | Allee der Kosmonauten |
| 15 | 20 | Anniversary Retreat |
| 16 | 19 | Grand-pre |
| 17 | 19 | Joyland |
| 18 | 19 | Pepito y los robachicos |
| 19 | 19 | Black Friday |
| 20 | 19 | Al Al Carnaval |
| 21 | 18 | Botate asobi |
| 22 | 16 | ’Huff’: Around the Edges |
| 23 | 15 | Here’s Television |
| 24 | 15 | Erbe wird gesucht, Ein |
| 25 | 14 | Chuji tabi nikki: Shinshu kessho hen |
| 26 | 14 | Hakob Hovnatanyan |
| 27 | 14 | Samho talchul |
| 28 | 13 | Du hao |
| 29 | 13 | Einflle der heiligen Klara, Die |
| 30 | 13 | Going Places with Lowell Thomas, #1 |
| 31 | 12 | Pitanje |
| 32 | 12 | Fuji ni tatsu kage |
| 33 | 12 | Dzhoy i Druzhok |
| 34 | 11 | Bar-L Ranch |
| 35 | 11 | Kalkofes Mattscheibe Sylvester Spezial |
| 36 | 10 | Geulim ilgi |
| 37 | 10 | Roof to Roof |
| 38 | 10 | Brick Wall |
| 39 | 10 | Dil Ki Duniya |
| 40 | 10 | Alte Snder, Der |
| 41 | 10 | Buddy Holly Story, The |
| 42 | 10 | Kun Hunttalan Matti Suomen osti |
| 43 | 9 | Sekret Enigmy |

For example we selected island 6 – ’Dona Macabra’.

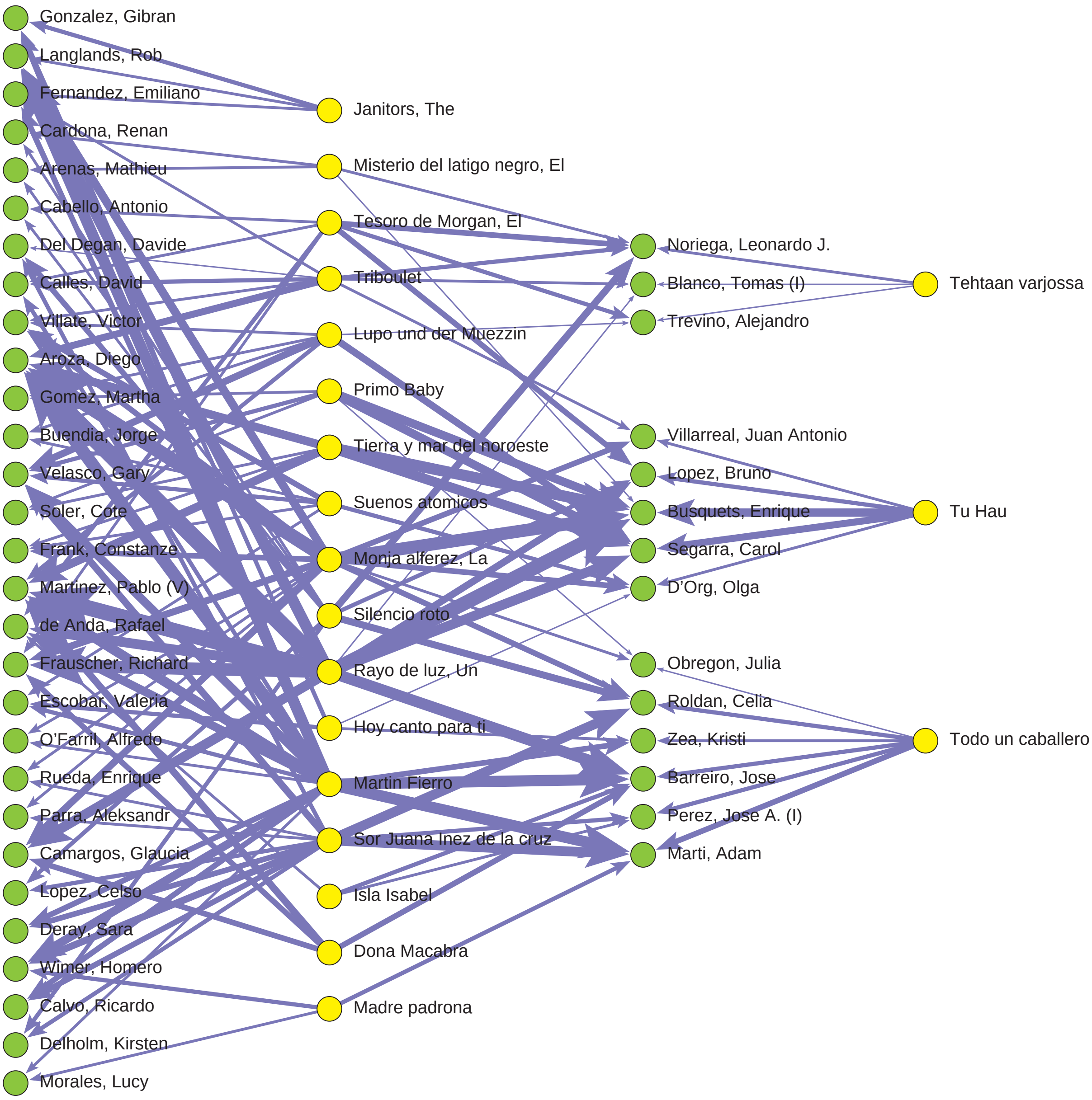



Figure 10: *Dona Macabra*

## Co-starring authors

We extracted a small subset of the actors in the IMDB network and constructed from it a dynamic visualisation of a 1-mode network showing the co-appearance of actors in films. This visualisation forms the first section of an animation, downloadable from the following location:

http://www.it.usyd.edu.au/∼dmerrick/gd05contest/gd05-final.avi

To define a sufficiently small subgraph, we first considered only nodes in the network with a Kevin Bacon number of 1. The Kevin Bacon number of an actor is a similar concept to the Erdös number of a mathematician; it represents the length of the shortest path in the movie star collaboration network from the actor to Kevin Bacon.

The data set was divided into time slices of a decade in length (e.g. 1920s, 1930s, etc.), and the set of actors reduced in each decade to only those who had co-starred in at least 5 films with another actor with a Kevin Bacon number of 1.

The 1-mode co-starring networks of these reduced sets of actors were constructed for each decade, and a three-dimensional force-directed layout generated for each. Nodes in the force-directed layout were restricted to lie on one of three concentric spheres, depending on the degree of the node, as illustrated in Figure 11. The colouring of each node was also used to indicate the degree. The size of each node was dependant on the number of movies in which the corresponding actor starred in that particular decade. Similarly, the width of an edge was used to represent the number of co-appearances between two actors in a decade.

To effectively illustrate the evolution of the co-starring network, we display smooth animations between the layouts of subsequent decades. The animations are broken into several parts shown one after the other in time, in order to aid retention of the mental map. First, nodes and edges not present in the first layout are faded out. Nodes present in both first and second layouts are then animated to their new positions in the second layout. Nodes new to the second layout burst out from the centre and come to rest in their calculated positions, and finally new edges are faded in to show the new collaborations in the second decade.

This process was continued for all decade slices from 1911 through to 2004, and the result can be seen in the downloadable animation.

The visualisation shows both expected and unexpected patterns. For example, nodes corresponding to singers Britney Spears, Beyoncé Knowles and Jennifer Lopez are highly connected, presumably due to music videos and attendance in music industry award ceremonies. Names of US presidents can be seen amongst a highly-connected component in the later decades, showing the wide-ranging scope of the genres in the IMDB. At one stage this component of political entities is seen to be linked to the network of movie stars through actor-cum-governor Arnold Schwarzenegger.

A more unexpected finding was the substantial number of actors with a Kevin Bacon number of 1 in the early years of the twentieth century, some of whom could clearly not have co-starred in a film with Kevin Bacon. This revealed some noise in the original contest data set. The years of some movies had been recorded incorrectly, while edges to other movies that possessed the same name as a movie of a prior decade were all recorded as belonging to the earlier movie.

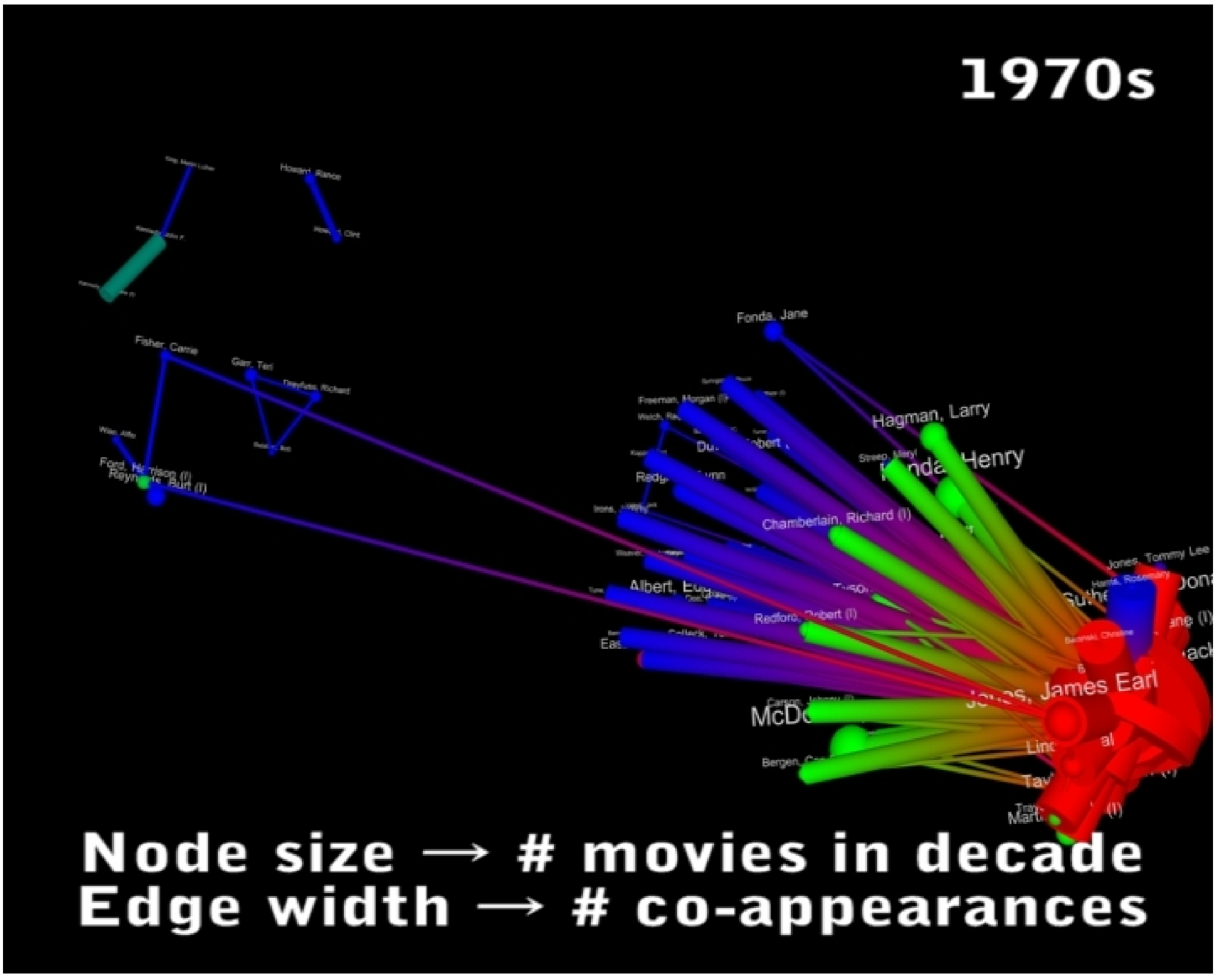



Figure 11: *A frame from the co-starring actors animation*

## A Galaxy of Movie Stars

Our final visualisation consists of a ”galaxy of stars” metaphor for the movie-actor network, and forms the second part of the animation downloadable from:

http://www.it.usyd.edu.au/ dmerrick/gd05contest/gd05-final.avi

A subset of the IMDB was selected for each year from 1907 to 2004. Actors and movies were chosen using the following criteria:

- every actor must have starred in more than 12 movies over the whole time period
- every movie must have more than 12 actors
- each actor must have played in between 3 to 6 movies in each year

A two-dimensional force-directed layout was generated for each year's subgraph. In the final visualisation, actor nodes in the network were depicted as stars in the night sky, and edges as faint lines joining up ”constellations” of actors (See Figure 12). Edges are present between actor and movie nodes, but movie nodes are hidden; in this manner, collaboration between actors can be seen. Animation is performed between each layout, in a similar manner to the animation of the co-starring authors network (detailed in the previous section).

No labels are shown in this visualisation, but the changing frequencies of highly-connected components can be seen as the visualisation changes over time.

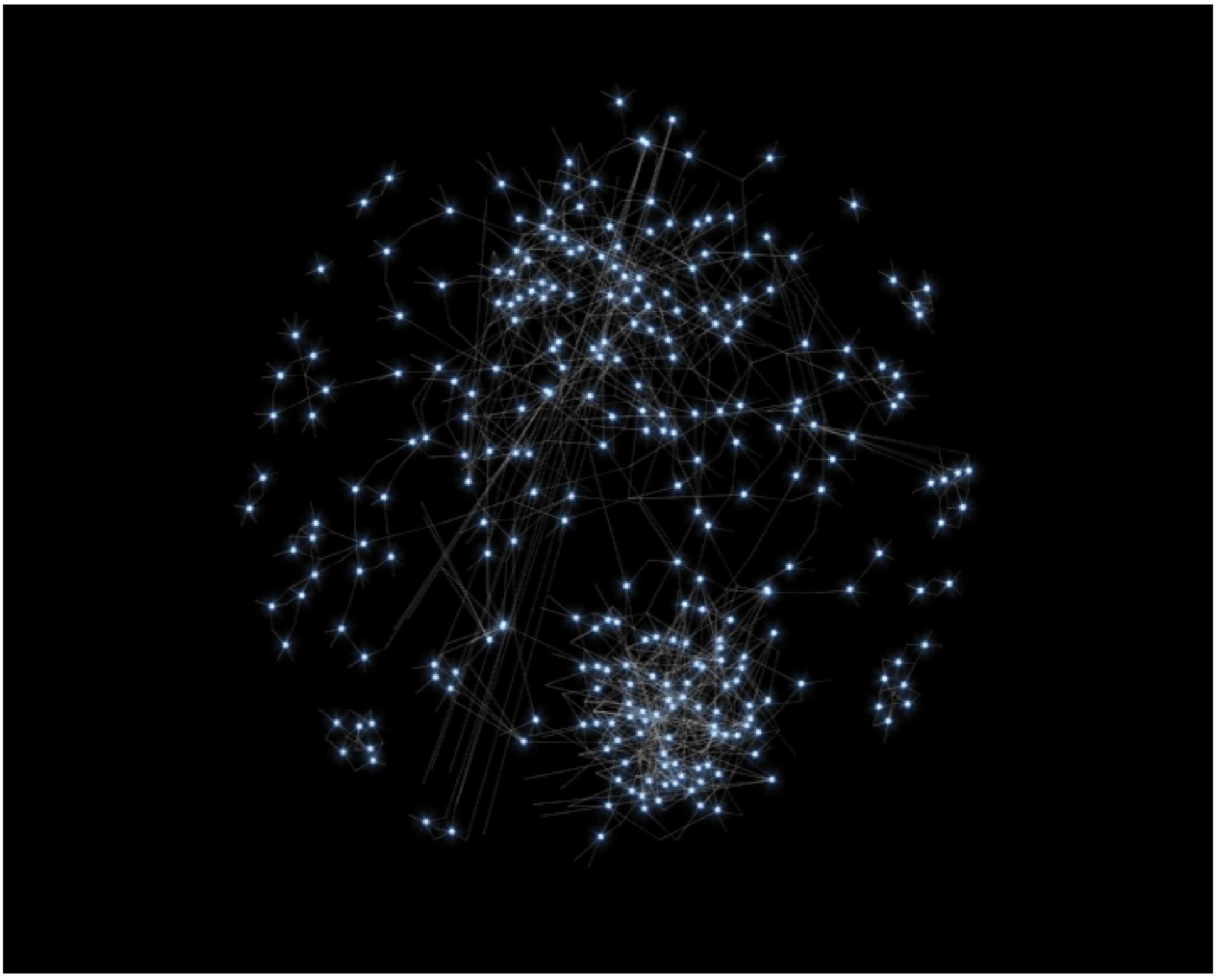


Figure 12: *A frame from the galaxy of stars animation*

# References


[1] Batagelj, V. and Mrvar, A.(1996-): ***Pajek*** *– program for analysis and visualization of large network*, home page, data sets.

[2] Batagelj, V. and Zaveršnik, M.(2002): *Generalized Cores*, arxiv cs.DS/0202039

[3] Batagelj, V. and Zaveršnik, M.(2003): *Short cycles connectivity*. arxiv cs.DS/0308011

[4] de Nooy, W., Mrvar, A. and Batagelj V. (2005): *Exploratory Social Network Analysis with Pajek*, CUP. Amazon. ESNA page.

[5] Zaveršnik, M. and Batagelj, V. (2004): *Islands*. Slides from *Sunbelt XXIV, Portorož, Slovenia, 12.-16. May 2004*, PDF