Movies and Actors: Mapping the Internet Movie Database

Bruce W. Herr, Weimao Ke, Elisha Hardy & Katy Börner School of Library and Information Science, Indiana University, Bloomington, IN 47405 {bherr@indiana.edu, wke@indiana.edu, efhardy@indiana.edu, katy@indiana.edu}

Abstract

This paper presents the results of an analysis and visualization of 428,440 movies from the Internet Movie Database (IMDb) provided for the Graph Drawing 2005 contest. Simple statistics are presented as well as a tapestry of all movies with an overlay of the giant component of the co-actor network. Academy award winners are highlighted. Major insights are discussed.

Keywords---network analysis, domain visualization, movies

1. Introduction

Since 2002, the International Sunbelt Social Network Conference has hosted a so called Viszards session [6] that aims to show the power of network analysis and visualization. The work discussed in this paper was done for Viszards 2006 at Sunbelt XXVI which took place in Vancouver, BC, Canada on April 28th, 2006. Viszards 2006 asked network science researchers to analyze data retrieved from the Internet Movie Database (IMDb). IMDb (<u>http://www.imdb.com</u>) is a popular site cataloging almost every movie ever made.

The study of IMDb data is interesting for several reasons. For one, most people know about and can relate to movies and actors. Thus, when presented with a visualization of movie data, they will try to find their favorite movies and actors, identify movies of potential interest or explore the complex co-actor relationships among actors. Second, the dataset has rich information on each movie and actor allowing for a wide variety of data analyses. Third, the dataset is sufficiently clean and structured so that analysis can be done without using semantic matching techniques.

From the beginning, our goal was to show all movies as well as major co-actor relationships. We wanted to give the world an overview of the movie and actor space that almost everyone is familiar with. Doing this on a large canvas (the final visualization has a size of 36" high and 73" wide) and in a way that people can

reason about and understand the visualization was a major challenge. The required data density due to data volume per square inch posed additional difficulties.

With this paper and the IMDb visualization we hope to communicate the power of visually pleasing yet informative visualizations to a general audience. Visualizations can be more than eye candy. Paper printouts are discussed as a viable alternative for the presentation of high density visualizations.

The remainder of the paper is organized as follows: Section 2 introduces the dataset used. Section 3 explains the data analyses and results. Section 4 discusses the iterative design of the visualization and insights gained. The paper concludes with a discussion and outlook.

2. Data preparation

The data for the IMDb visualization originates from the Graph Drawing 2005 web site [3] at <u>http://www.ul.ie/gd2005/dataset.html</u>. The dataset is a bipartite graph in which each node either corresponds to an actor or to a movie. Edges go from a movie to each actor in the movie. It also provides metadata for the nodes like movie/actor name, year of the movie, and genre of the movie. This data was then parsed and stored in a relational database to ease data manipulation.

As with all large datasets, there were diverse anomalies. Out of the 428,440 movies in the set, 2,091 movies had no year data, six movies were produced in 1 CE, two were produced in 2 CE, 24 more were produced between the years 3 and 1888 CE, and the 'Adult' movie entitled 'Westside Boys' is to be produced in 9006 CE. The biggest anomaly in the derived data is the fact that of the 428,440 movies provided, 123,617 movies have no actor data at all. This is particularly problematic for us since we are showing the interplay between actors and movies. We believe that this is most likely a problem inherited from the derived data, since the official IMDb statistics say that as of March 2007 (the derived data was from early 2005) there are 365,328 movies in the database. In the end, we excluded those movies that did not have actor information.

3. Data analysis

After getting the data into a relational database, several statistics were run to get a feel for the data. We excluded the anomalous data discussed in the last section resulting in 302,691 movies produced between 1890 and 2007. It should be noted that this data was from early 2005, so all movies beyond that were in differing stages of production and not yet released. Figure 1 shows the growth of movies over time. The red lines show the boundaries of the movies that we considered.



Figure 1. Growth of Movies Over Time



Figure 2. Movie Out-Degree Distribution

There are 3,792,390 links connecting 302,691 movies and their 896,308 actors, see Figure 2. 38,027 movies have exactly one actor. Movies with more than 1,000 actors are 'The Eurovision Song Contest' (1,338 actors), 'Around the World in 80 days' (1,287 actors), and 'General Hospital' (1,123 actors).

In order to get a feel for the actor space, we created a co-actor network where actors are connected based on the movies they acted together in. Actors that appear in a movie together are said to co-act. The network of coacting actors contains 896,308 actor nodes and 114,128,535 co-actor links, see Figure 3. Each link is weighted by the number of movies the two actors were in together.



Figure 3. Co-Actor Out-Degree Distribution

4. Data visualization

Our main goal with this visualization was to give a global overview of the entire movie and actor space. During the initial design phase, we wanted to draw a coactor network and have it be surrounded by a list of all movies. However, fitting this into a reasonably sized canvas proved difficult. The sheer number of movies to plot was so large that we eventually decided to render them in columns and overlay the co-actor network directly on top of the movies. We also wanted to constrain ourselves to 36" high and somewhere around 40" wide, but eventually went with 73" wide to accommodate all of the movies. A description of the layers of the final visualization follows.

At the bottom of the visualization is the movies layer. The movies were grouped by year and plotted in 97 columns. Within each year, the movies were sorted and their titles size encoded by the number of starring actors. Furthermore, movie titles are sorted and color coded by genre. Each of the seven top genres (Short, Drama, Comedy, Documentary, Adult, Romance, and Thriller) was given a distinct color while the rest were given a light grey color. Over plotting was utilized to fit the movies into the area provided. A white outline is drawn around each character to improve text legibility. A close-up of the movies layer is given in Figure 4.



The next layer up is the actor layer. We felt that the best way of showing the actors was by laying out their co-actor network using a force-directed layout algorithm. Each edge between the actors was weighted by the number of times the two actors had been in the same movie together. Interested to see the strongest co-

actorship linkages we excluded all those links that had a weight of less than three. Resulting unconnected nodes were excluded. The remaining core network with 105,758 nodes and 1,292,816 edges was fed into VxOrd [2] to lay out actor nodes with a modified spring force layout algorithm. This algorithm ensures that highly interlinked nodes are close to each other and unlinked or weakly linked nodes are further apart. The resulting list of coordinates for each of the 105,758 actor nodes was rendered using Pajek [4]. The color of each actor node corresponds to the movie genre s/he most contributed to. The results are shown below in Figure 5. A zoomed in portion of the co-actor network can be seen in Figure 6.



Figure 5. Co-Actor Network



Figure 6. Zoomed View of the Co-Actor Network

Another layer was added to provide landmarks in this complex co-actor network. The network was spatially cut into a 10x10 grid and the actor node that had been in the most movies in each of the cells was labeled with their actor name using a light colored, 15-point font. These labels are useful in identifying clusters of actors.

The discussed layers form a reference system that can be used to overlay additional data. In the visualization described in this paper, we added two more data layers. The first shows Academy Award's best actor and actress winners and nominees from 2000-2004 [1]. They are represented as 41 darker and larger actor labels on top of the co-actor network. The most interesting part about this layer is that most of the actors are tightly packed into one cluster. Though not fully explored, this may mean that in order to increase one's chances of an academy award for best actor/actress, one should work closely with actors in this cluster. The second additional layer dealt with the winners and nominees for the Academy Award's best picture award. The 25 movies nominated (including the winners) have exactly 433 actors in the co-actor network. This layer draws lines from the 25 nominated movies in the underlying movie layer to the associated actors in the coactor network layer. The color of the lines corresponds to the genre of the movie. The curves of the lines were chosen so as to not cover up too much of the co-actor network. This layer helps to highlight what areas of the actor space is being used by top movies in the field.

All of the layers except for the co-actor network were created with custom code that reads in the provided data and produces PostScript® files. The co-actor network's layer was outputted to PostScript through the Pajek program. To produce the final image, the assorted layers were combined and rasterized at 400 dots per inch (DPI) in Adobe Photoshop©. An additional layer was created in Photoshop that added the informational column on the right side of the visualization.

The movies layer proved to be very difficult to rasterize in Photoshop due to its size and complexity. For the version presented at Sunbelt, we had to reduce its complexity by removing the textual outline drawing. This worked, but we were never quite satisfied with the loss in quality that resulted. After nine months of trying larger machines and distributed rendering, a solution was found. By using the GNU Image Manipulation Program (GIMP) and utilizing a Sun server with 32 GB of RAM and 4 processing cores, we finally got the layer to satisfactorily render at 400 DPI. This new image has replaced the older movies layer.

Figure 7 at the end of this paper shows the final visualization. Unfortunately, it is more than eight times smaller than the original visualization and many details are lost at this size. To really appreciate the visualization, one must either have a full resolution printed version or go to <u>http://scimaps.org/maps/movieactors</u> to see a zoomable Google Maps interface to the visualization. The map is also available for sale from <u>http://scimaps.org/ordermaps</u> in support of the *Places & Spaces: Mapping Science* exhibit.

Discussion

The presented work demonstrates the utility of paper printouts for serving high data density visualizations. Paper as a medium is easy to access and transport, offers high data density, and is comparatively cheap. Humans have used paper and interacted with it for well over 2,000 years and have highly optimized it as a medium to store, transmit, and preserve information. Paper naturally supports exploration. Interactivity like zooming and panning can be accomplished by physically moving closer to and further away from the print. While there are problems with zooming and panning in computational environments, this sort of interaction with paper is immediately obvious to viewers. Arbitrary annotations are possible. Last but not least, there is something to be

said about a visualization that can be physically touched and has a real texture to it.

The higher density of paper has allowed us to give an overview of the entire movie and actor space in a reasonable physical space. Our current visualization renders at 400 DPI, but there are techniques to utilize up to 4,000 DPI. The result is extremely crisp graphics that allow for further zooming with a physical magnifying glass.

In addition to bringing out paper's natural strength, this visualization work also made obvious the current limitations of rendering on large display walls. Display walls are limited by the rather low resolution of modern monitors and projectors. To render the full resolution visualization, a display wall would have to be around 4 times as large as the equivalent print. A 12'x24' display wall would be prohibitively expensive. Compare this to an equivalent 3'x 6' print which is much cheaper, denser, and could be mass produced.

Future work aims to update the data behind the visualization and add a layer of interactivity. We will do this by utilizing an invention by W. Bradford Paley called an illuminated diagram [5]. This technique uses a projector to interactively highlight interesting parts on statically printed diagrams. We can then take advantage of the interactivity of computers, yet still retain the qualities of printed media.

Acknowledgements

We would like to thank all those involved in the Internet Movie Database for creating an excellent dataset, Vladimir Batagelj for organizing the Viszards session, Bryan J. Hook for editing, and Sumeet Ambre for creating the Google Maps interface for our visualization.

This research is supported by the National Science Foundation under IIS-0513650 and a CAREER grant under IIS-0238261. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

- Academy Award's best actor and actress winners and nominees from 2000-2004 downloaded from <u>http://www.imdb.com/Sections/Awards</u>. Accessed on April 2006.
- [2] Davidson, G.S., Wylie, B.N. and Boyack, K.W. Cluster stability and the use of noise in interpretation of clustering. *Proc. IEEE Information Visualization 2001*. 23-30.
- [3] Internet Movie Database (IMDb) network provided for GD'05 at <u>http://www.ul.ie/gd2005</u>.
- [4] Nooy, W.d., Mrvar, A. and Batagelj, V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.
- [5] Paley, W. Bradford. Illuminated Diagrams: Using Light and Print to Comparative Advantage. *InfoVis Conference* 2002.

[6] Viszards: Analysis and Visualization of IMDB Networks session description from <u>http://www.insna.org/2006/special.sessions.html</u>. Accessed on April 2006.



