Movie Analytics: Visualization of the Co-starring Network

Dominique Haughton*
Bentley University,
Université Paris 1
Université Toulouse 1

Mark-David McLaughlin
Bentley University
Cisco Systems

Kevin Mentzer
Bentley University

Changan Zhang
Bentley University
Epsilon

ABSTRACT

This poster contributes a novel application of social network visualization techniques to the motion picture industry. We make the case and illustrate with examples that a visualization approach based on k-cores helps alleviate otherwise inextricable memory issues in analyses of the IMDb co-starring network, which contains more than 2.6 million actors displaying over a billion links, with degrees which can rise to about 50,000 and above for the most connected actors.

Keywords: Movie Analytics, k-core decomposition, IMDb.

Index Terms: G.2.2 [Graph Theory]: Network Problems; E.1 [Data Structures]: Graphs and Networks.

1. INTRODUCTION

Attempts at visualizing the co-starring network, in particular using the Internet Movie Database (IMDb, owned by Amazon) network, as well as various attempts at visualizing other aspects of movies, such as the network of actors featured in the same scene within a movie, story lines etc. abound in published and unpublished reports. We do not attempt here to extensively review this literature, but a simple search of keywords such as visualization, IMDb database, co-starring network yields a representative set of graphs and reports.

Interesting efforts (in the context of the VAST 2013 challenge) at using visualization tools to help predict box-office revenues and ratings include [1].

Network/Partition	Number of Actors	Number of Connections	Average Degree
Complete Network	2,614,790	1,083,901,049	834.05
Most Connected (30k+)	8,432	56,661,389	13,439.61
Minor Actors (10-50)	986,031	5,584,885	11.33

Table 1: Summary statistics for the IMDb database

To the best of our knowledge, no one has found a way of representing the entire network of about 2.6 million actors (see Table 1), in large part because of the difficulties with fitting the co-starring network of these actors with over a billion links into memory. With both the Pajek and k-core techniques used in this poster, attempting to visualize the entire co-starring network leads to memory errors such as segmentation faults or memory locks. Another challenge is that of making sense of visualizations given that identifying a particular actor in such vast representations is difficult as well.

We propose to discuss and demonstrate a k-core approach to this visualization and analysis challenge. We rely on past work and algorithms due to [2] and compare the visualizations we obtain via the k-core approach with those obtained by more traditional methods.

2. VISUALIZATIONS OF THE IMDB CO-STARRING NETWORK

We present here visualizations of two subsets of the co-starring network: that of the most connected actors (with degrees of 30,000 and above, in Figure 1, top panel) and that of the "minor actors" (with degrees between 10 and 50, in Figure 1, bottom panel). The graphs were produced with Pajek [3] [4].

Figure 1 (top panel) reveals the presence of clusters within the network, numbering about 7 at first sight, and Figure 1 (bottom panel) gives a sense of a diffuse network with many inter- and intra- connected smaller groups.

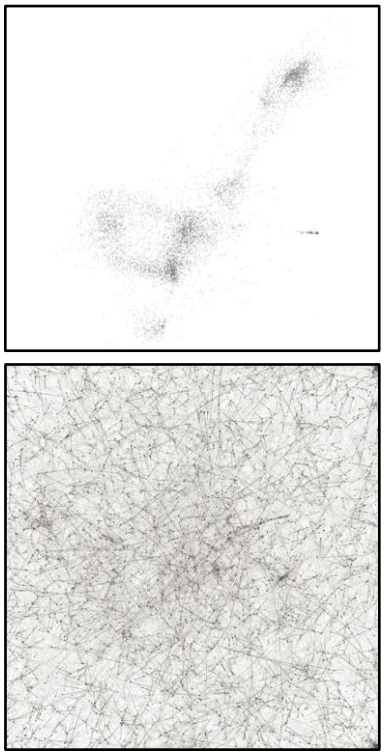

Figure 1: Representation of the network of most connected actors (top panel) and of the minor actor network (bottom panel) with Pajek [3] [4].

IEEE Symposium on Large Data Analysis and Visualization 2014
October 9–10, Paris, France
978-1-4799-5215-1/14/$31.00 ©2014 IEEE

*dhaughto@bentley.edu

K-core Representations of the IMDb Co-starring Network

We recall that [2]: 1. A sub-graph H of a graph C is a k-core if the degree of all vertices in H have degree in H greater than or equal to k and H is the maximal sub-graph with this property. 2. A vertex has shell index c if it belongs to a c-core but not a (c+1)-core. 3. A shell of order c is composed of all vertices whose shell index is c. 4. Each connected (within the original graph C) set of vertices with shell index c form a cluster of order c.

Figures 2 and 3 display a k-core representation of the network of the most connected actors and that of the network of minor actors respectively. We recall [2] that:

- The color of the nodes represents their shell index
- The size of the nodes represents their degree in the original network (in logarithmic scale)
- Nodes are displayed in concentric circles with a radius which is larger for vertices with a lower shell index
- Within a c-shell, c-clusters (connected - within the original graph - nodes within the same shell) are displayed near each other
- A random sample of edges are displayed; the two halves of each edge are colored with the color of the corresponding extremities
- For a given shell diameter, vertices can be placed more internally (if they are connected to vertices with higher shell indices) or externally (if they are connected to vertices with lower shell indices)

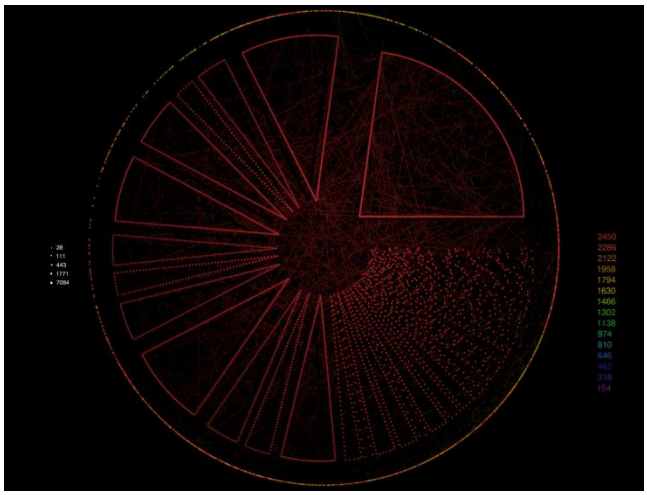

Figure 2: K-core representation of the network of most connected actors.

Figure 2 reveals the presence of about 13 main groups of actors as well as at least 20 more diffuse groups. Very few actors lie in the maximum shell in the very center of the graph. We note that the size of the nodes does not clearly increase as one moves from the outer to the inners circles. The "pie slice" structure of parts of the graphs is quite interesting. It indicates that an actor with a relatively high shell index links to a chain of connected actors from different shells, which in turns links to a cluster in an outer shell.

We note that the range of colors is quite narrow, indicating that most nodes lie within a relatively small portion of the full range of possible shell indices (from minimum to maximum).

We also note that the outer shell is made up of actors who mostly connect among themselves and not to actors with different shell indices.

Next, we examine and discuss a similar representation, but this time for the network of minor actors.

The k-core representation of the network of less connected actors (Figure 3) is strikingly different from that of the most connected actors (Figure 2). Numerous circles represent components within the k-shells which are disconnected from each other. The k-core approach allows for a clearer view (compared to that in Figure 1) of the many small communities in the network; this is made possible by the successive "peeling" of the k-cores as k increases.

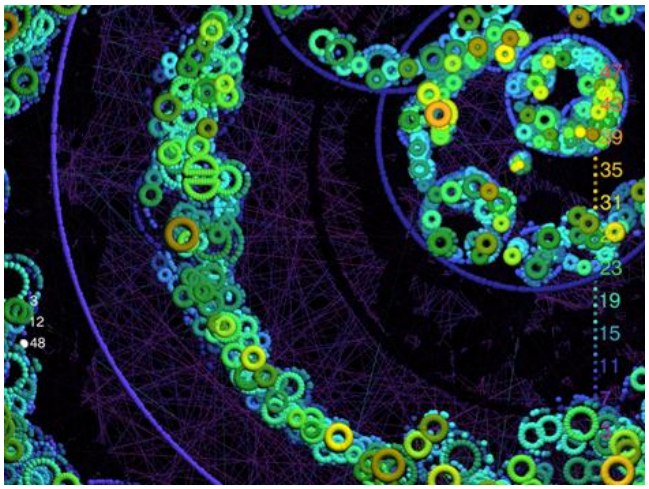

Figure 3: K-core representation of a portion of the minor actor network.

3. Conclusion and Future Perspectives

This poster has contributed a novel application of network visualization techniques to the area of movie analytics. Visualization is an important component used for understanding the dynamics of a network, but up to this point researchers have struggled with a way to visualize very-large datasets. This suggests to the researcher one technique that can be employed to overcome this limitation.

Further perspectives include investigations of k-core representations of other subsets of interest within the IMDb co-starring network, notably with an identification of key central actors and groupings by genre, director etc.

References

[1] M. el Assady, D. Hafner, M. Hund, A. Jager, W. Jentner, C. Rohrdantz, F. Fisher, S. Simon, T. Schreck and D. Keim. "Visual analytics for the prediction of movie rating and box office performance, http://bib.dbvis.de/uploadedFiles/elassady.pdf. 2013.

[2] J. Ignacio Alvarez-Hamelin, L. Dall'Asta, A. Barrat and A. Vespignani. "K-core decomposition: a tool for the visualization of large scale networks", http://arxiv.org/pdf/cs/0504107.pdf, 2005.

[3] W. de Nooy, A. Mrvar and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2011.

[4] Pajek. *Program for large network analysis*, available at http://pajek.imfm.si/doku.php?id=pajek, 2014.