Overlapper: visualization of interconnected social groups

Roberto Therón, Rodrigo Santamaría, Juan García, Diego Gómez, and Vadim Paz-Madrid

Departamento de Informática y Automática, Universidad de Salamanca

1 Introduction

Graph drawing has problems with scalability when the number of nodes and connections grow. As a result the performance is compromised and the edge representation becomes too cluttered. Filtering data is generally used to decrease graph complexity: previous filtering through external tools, neighbor criteria and focus+context techniques are some examples of filtering.

In order to reduce edge cluttering, different edge drawings (splines, transparency, hiding) have been developed with different degrees of success. Social networks can usually be separated in groups, such as close friends, people involved in the same project or families. Using this grouping, zone graphs have been developed, wrapping separate groups of nodes by areas.

We have developed a zone graph directed by forces regarding relationships in Hollywood movies. The use of *zones* avoids *edge cluttering*, while the use of *filters* (by means of ancillary linked views, such as parallel coordinates or scatter plots) used previously to graph drawing reduces *graph complexity*. Overlapping zones and representation of nodes help in finding interesting relationships within the graph, such as people working together in different movies, hub nodes or paths between nodes.

2 Methods

2.1 Group Definition

We consider a 'group' the whole cast for each movie. This concept can be directly transferred to other social networks as it is the case of authors of an article or workers involved in a project. In the case of movies, group's sizes range from 2 to more than 200 people. Data have been obtained from US movies made after 2000, collecting the corresponding information from www.imdb.com.

Each movie M_k is displayed using an undirected complete graph where nodes are the set of people involved in the movie, $n_1, ..., n_{p_k}$. Edges connect nodes that appear in the same group but, to avoid edge cluttering, edges are not drawn and instead semi-transparent hulls wrap each group (eq. 1).

$$E = \{(n_i, n_j) / \ \exists\, M_k \ with \ n_i, n_j \in M_k\} \tag{1}$$

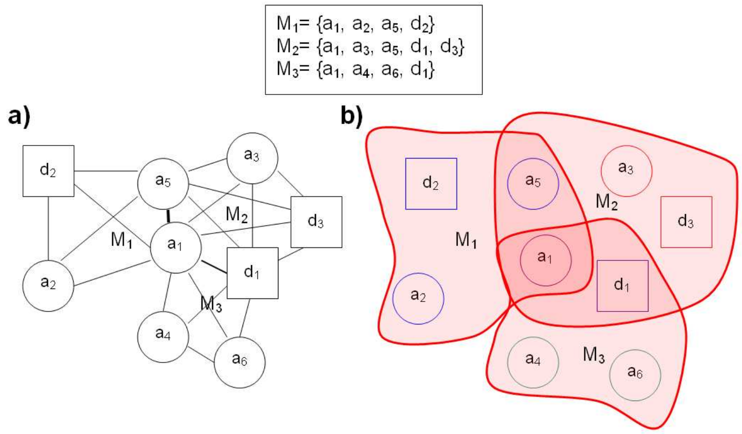

Fig. 1. Three movies and their representation. Different roles in the movie are represented by different shapes (here, for example, directors are squares and actors are circles). d_1 and a_1 appear on both M_1 and M_3 so the edge between them is shorter (a) and they appear in the intersecting area in the final visualisation (b). Actor a_1 is also in M_2. Intersecting areas are more opaque to highlight overlapping.

Therefore, each movie $M_k = (n_1, ..., n_{p_k})$ is a complete subgraph of p_k nodes (Fig. 1). The weight of an edge $e = (n_i, n_j)$ is given by the number of movies that contain both persons n_i and n_j.

Unlike other zone graph visualisations as those of [4] or [3], a node can be in more than one zone, reflecting overlapping between movies, that can usually affect to more than a node. Hulls are drawn with a transparent color, so intersecting areas become more opaque.

2.2 Graph Layout

The nodes are displayed following a force-directed layout [2]. In our model, each pair of nodes can be affected by up to two forces. If the nodes are connected, a spring force acts to keep them at an optimal distance, with a determinate stiffness. Between every pair of nodes, connected or not, an expansion force makes them to repel each other. The string force keeps nodes in the same movie close, while the expansion force separates nodes in different movies. Fig. 2 illustrates the entire process of graph building and layout.

2.3 Node Representation

Node positions are defined by the graph layout, but additional information is present at the level of node representation, by the means of glyphs, at user's demand. A glyph is a graphical object designed to convey multiple data values [5]. The size, color and shape of the glyph represent different dimensions.

In our case:

```
N=∅: set of nodes (persons)
E=∅: set of edges
G: expansion constant
s: stiffness constant
d_o: optimal string length
Construction phase:
For (each movie M_k)
    For (each node n_i in M_k)
        If(n_i not in N)
            add node n_i to N
            assign random position (x_i,y_i) to n_i
            For (each object n_j in M_k, n_j≠n_i)
                If(edge (n_i,n_j) not in E)
                    add edge (n_i,n_j) to E
Layout phase:
Do
  For(each node n_i in N)
    For (each node n_j in N, n_j≠n_i)
      d_ij= Euclidean distance (n_i,n_j)
      apply expansion force X_ij= −G/d_ij
      If(edge (n_i,n_j) in E)
        apply string force S_ij=s(d_ij-d_o)
Until stability
```

Fig. 2. Graph drawing algorithm.

- The shape of each node distinguishes between roles. Shapes are designed in a way than overlapping of roles can be represented and easily identified (see fig 3).
- Pie charts with as much sectors as movies in which the node appears can be superposed to node shapes (color of sectors could also be used to identify different movie subgroups following some criterium).
- Labels with gene and condition names can be displayed to help with node identifying. In this case, label color is determined by the person gender and text size by the number of films in which the person works.
- A brief description and a photograph of the person is recovered on demand from the internet and drawn next to the corresponding node.

2.4 Graph Interaction

In order to foster knowledge discovery, the visualisation is not an static image, but a user driven representation that can be manipulated in a number of ways. Besides controlling node drawing explained above, user can:

- Change force parameters G, d_o and s.

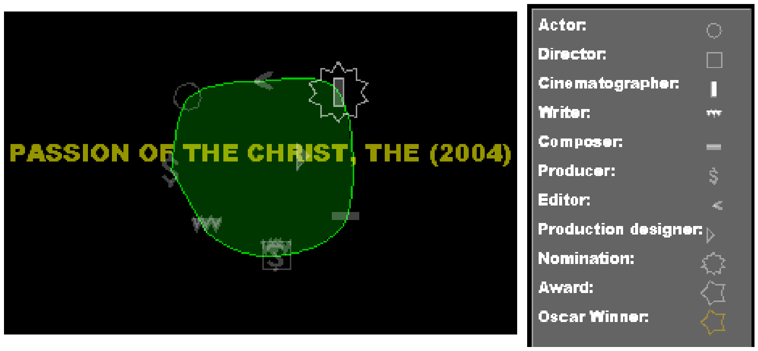

Fig. 3. Roles legend and it's application in a movie. We can easily determine that the cinematographer of 'The Passion of the Christ' have been nominated, there are two writers, one of them being the director and co-producer of the movie.

– Modify the representation by dragging nodes and by fixing node positions.
– Search for person names or film titles.
– Visualise or hide edges and hulls.
– Highlight the nodes connected to a particular node.
– Overview and detail of the represented graphs.

The final result of the graph display is a set of flexible overlapped, colored areas representing movies, with glyph nodes inside representing persons and their different roles. Drawing these areas instead of drawing edges, along with its flexibility, allows a large number of films to be represented without excessive cluttering on the display. Fig. 4 shows how the layout works with a simple case, giving insight of how three movies relate with just a glance.

2.5 Graph Complexity

The most important drawback of force-directed layouts is complexity. An optimal implementation has a complexity of $O(n^3)$ [1], being n the number of nodes. The drawing of the whole set of movies (more than 20.000, with more than 300.000 persons involved) is obviously impossible and not very useful anyway.

To reduce this dimensionality, filters have been implemented. First, a movie filter through different simple visualizations (scatter plots and parallel coordinates) allow to select films under numerical criteria (box office earnings, number of awards, year, imdb ratings, etc.) Second, a text filter allow to select only movies with a determinate title, genre or person involved. Finally, the cast of each movie can also be filtered by any threshold to only visualize the most relevant people involved.

Playing with filters, our tool can deal without relevant loss of interactivity with up to 500 nodes, on an Intel Pentium D 2.8 GHz processor. This allows analysis of the 50 most profitable movies, comparisons between the top-ten movies by budget and the oscarized movies or the inspection of an actor's filmography, for example.

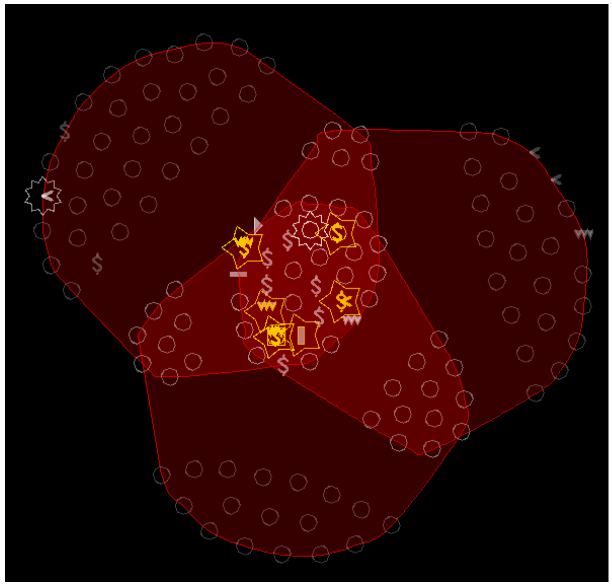

Fig. 4. Visualization of the trilogy of 'The Lord of the Rings'. Zones for people involved in one, two or the whole trilogy are easily distinguishable. Deep insight is achieved with just a glance. For example, occasional involvements are usually actors and the persons that won awards because of the trilogy were present in the all its films.

2.6 Graph Overlapping

Overlapping is usually seen as a problem in zone graph drawing because overlapped zones hide information of objects under it, it is difficult to identify the group to which a node pertains if it is in an overlapped area and eventually zone cluttering could appear as happens with edge cluttering.

We have tried to make overlapping a part of the solution instead of part of the problem by the use of different techniques. In first place, the use of transparency to fill zones (keeping an opaque border to identify limits) makes easy to identify overlapping areas, avoiding to hide nodes or areas under it. Fig.4 shows how areas with one, two or three overlapping zones are easily identifiable thanks to transparency degrees, without losing the identity of each movie.

On the other hand, when a node is in a area corresponding to two overlapping zones, pie chart representation of nodes helps to discern if the node is in both areas (two sectors) or only one (no pie chart). Due to force directed layout, most of the nodes will appear well positioned when stability is reached, but until then, or in the case of very connected graphs, pie charts can be very helpful.

Finally, apart from the overlapping of zones of different movies because of intersecting persons, our implementation allows the use of color to identify movies under different numerical or textual criteria, with the use of filters. This way we can, for example, visualize Steven Spielberg's movies and awarded movies with different colors (for

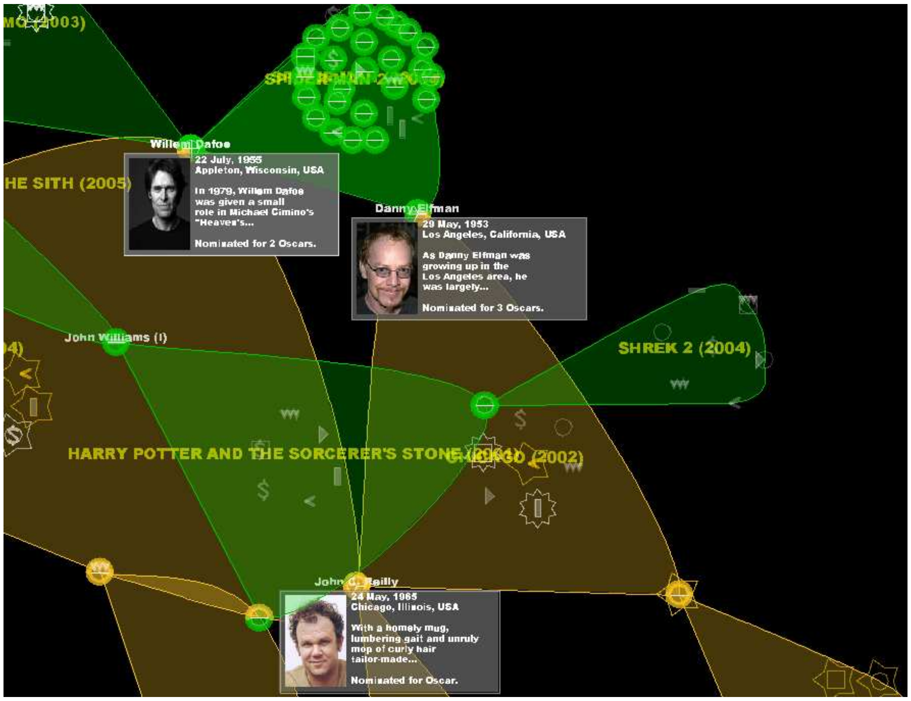

Fig. 5. Detail of the visualization of the most awarded movies (yellow) and the films that earned more money (green). We can observe how overlapping and pie charts helps us with graph interpretation without effort. Top group of two-sector pie charts with more opaque zone are the overlapping of the two movies of Spider-Man, with nearly the same cast. John Williams (center-left), although inside a yellow movie, is easily distinguished as out of the movie because of the pie chart and the green zones connections. Danny Elfman is in three movies (three sectors, two green and one yellow), participating in the Spider-man saga and in an awarded movie, 'Chicago'. By contrast, John C. Reilly appears in awarded movies but not blockbusters, working each time with completely different sets of people.

example, yellow and cyan) and see how these sets of movies interact (awarded movies of Steven Spielberg will appear as green zones, with blue/yellow piecharts for all the people involved) Fig. 5 shows this and other features discussed above.

3 Credits

This work has been done by the VisUsal Group of the University of Salamanca. A web page including this resume along with lots of snapshots and video demos of how Overlapper works is available at http://carpex.usal.es/∼visusal/site.

References

1. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, /2000.
2. J. B. Kruskal and J. B. Seery. Designing network diagrams. *Proceedings on the First General Conference on Social Gaphics, Washington D.C.*, 1980.
3. G. Kumar and M. Garland. Visual exploration of complex time-varying graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):805–812, September/October 2006.
4. A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions onf Visualization and Computer Graphics*, 12(5):693–700, September/October 2006.
5. C. Ware. *Information Visualization: Perception for Design*. 1999.