Case Study of Co-authorship Networks Using a Tool for Graph Visualization

Andre Suslik Spritzer, Fernando Profs Volquind, Carla M. D. S. Freitas

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{spritzer,fpvolquind,carla}@inf.ufrgs.br

Abstract. *One of the most common applications of graph visualization is in the study of social networks. This paper shows how a paper co-authorship network can be explored using a graph visualization technique. The technique utilized builds upon the physics metaphor of traditional force-directed graph layouts to provide the user with interactive tools for the manipulation of the graph. These tools consist basically of magnets, which can be set to attract nodes and edges depending on their topological and semantic properties. In this work we show how they can be simply used to answer common questions about co-authorship networks.*

1. Introduction

In the past few years the study of social networks has become especially relevant due to the growing popularity of applications that make use of them. In websites such as Facebook, MySpace, Orkut and Flickr, users can share, organize and find information as well as establish contacts depending on common interests. The analysis of social networks is necessary in many fields, including social and behavioral sciences, economy and marketing [Wasserman and Faust 1994]. While this can be done using statistical studies and other procedures that do not have a visual output, it is often the case that producing a visual representation of the network might make the same information more accessible and reveal new traits and relationships that were previously undetectable.

Since a social network is in essence a set of people and the relationships between them, the natural data structure to represent it is a graph. While information organized in graph-like structures can be usually explored textually, through tools such as query languages or the computation of metrics, this usually requires an expert user, being too complex and not intuitive enough for most people, who have little experience with such concepts. Therefore, many areas where graph structures apply, including social network applications, often require visualization techniques.

By far, the most popular and intuitive visual representation of a graph is the node-link diagram. Different algorithms for such layout have been created, with each favoring certain aesthetic criteria. Some of these techniques are better for certain applications while some are better for others, but all have their limitations, which range from computational cost to visual clutter. To deal with the limitations of these layout algorithms, many approaches have been experimented. While some have applied navigation and interaction schemes on the traditional layouts [Abello and Korn 2002], others have built alternative visualization techniques [Heer and Boyd 2005], or combined existent techniques into new hybrid ones [Henry et al. 2007b].

When building a layout, most graph visualization techniques do not take into account the potentially relevant information contained in the attributes of the nodes and edges, focusing only on the topology of the graph. Even though the main information associated to a graph is the relationships represented by its topology, it is often the case that the attributes represented in its nodes and edges are just as important to the user of a graph visualization application. This user might thus be missing out important data and even relationships that are not explicitly expressed by the structure of the graph.

In social network analysis, attributes of nodes and edges are very important and thus should be considered for layout computation. In this work, we describe how to use a physics-based graph visualization technique to interactively manipulate the drawing of a graph in order to make it semantically more relevant for answering typical questions in a social network application.

The paper is organized as follows. Next section briefly describes related work. Section 3 presents the visualization tool we have used in the case study described in Section 4. Section 5 draws some conclusions and outlines the future work.

2. Related Work

Graph drawing is an old and deeply explored subject, but due to recent applications dealing with large graphs, the interest has been directed to visualization techniques with interaction and navigation features to allow exploring static, pre-computed layouts. Well-known techniques include filtering; fish-eye views; scrolling and panning; zooming and even coordination of two or more visualizations [Herman et al. 2000]. Very few techniques, though, allow for dynamic interactive reorganization of graph layouts, as a means of facilitating relationships and attribute investigation.

In some cases the techniques allow for simple layout reorganization by letting the user move around nodes in force-directed layouts, which will trigger a repositioning of the nodes. Another known technique is to find clusters of nodes and transform them into cluster-nodes that can be expanded and collapsed by the user. Clusters can also be used to perform cluster-based semantic zooming, which allows for a level-of-detail-like approach to the visualization, letting the user incrementally explore the graph by zooming in or out.

Amongst the few techniques that allow for dynamic graph layout reorganization, we find NodeTrix [Henry et al. 2007a], which is a hybrid of matrix and node-link visualizations. NodeTrix allows the user to turn clusters of nodes of node-link visualizations into matrices, which are then displayed with the node-link diagram. Nodetrix was applied to visually explore graphs containing information from four human-computer interaction conferences [Henry et al. 2007b].

Another relevant work is Perer and Shneiderman's (2008) which used four long-term case studies to prove the power of integrating statistics and visual exploratory features in *SocialAction*, a tool for social network analysis. Employing this tool, four domain experts explored datasets making discoveries that would not be possible without the integration of statistics and visualization.

3. Visualization tool

This work uses the physics-based technique proposed by Spritzer and Freitas (2008) to provide users with means for interactively reorganizing imperfect graph layouts into ones that better fit their needs. This technique builds on the physics metaphor of force-directed graph layout algorithms to allow them to place virtual magnets in the scene. These can be set to attract all the nodes and edges that fulfill a given set of criteria, which can be topological or attribute-based. The prototype we developed to illustrate the technique allows the user to open a graph in GraphML format and explore it using interactive tools. The following subsections summarize the prototype's main features.

3.1. Layout

For the basic graph layout, Fruchterman and Reingold's technique [Fruchterman and Reingold 1991] is combined with the Barnes and Hut (1986) algorithm [Tunkelang 1999], and slightly adapted to better fit our needs. These modifications were the addition of a small gravitational force that pulls all nodes slightly towards the centre of the workspace and the alteration of the manner in which the algorithm runs (we re-evaluate it every frame instead of running it for a given number of iterations). The users are allowed to pause and play the physics simulation whenever they desire.

Besides the central gravitational force, the user can insert gravity spots in the scene to attract nodes to a certain region. A gravity spot can be set to affect either all nodes of the scene or only the nodes within a certain radius of its location. In the latter case, the user can set to display a circle with the desired radius.

3.2. Magnets

Magnets are special objects that can be added to the scene which have the ability to attract nodes of a graph that fulfils certain user-defined criteria. A magnet works by exerting onto each of these nodes an attraction force that will progressively move them towards it, thereby building a cluster of semantically-related nodes around it. When these nodes move, the force-directed layout algorithm ensures that all the other nodes that are connected to them by edges will be pulled along, reorganizing the whole layout of the graph in the process. A magnet that attracts all nodes that have a degree of 3 can be seen in Figure 1.

To each magnet users should associate one or more attraction criteria, which can be set as requirements of attraction or simply criteria. To be attracted a node must fulfill all requirements or at least one of the defined criteria. These requirements and criteria can be based on the topology of the graph, on attributes of its nodes and edges or even on other magnets that have been placed in the scene.

Topology-based criteria use the structure of the graph to attract nodes. It is possible to attract nodes based on properties such as degree, path length (all nodes that are within a specified path length from another node or group of nodes), connected subgraph (subgraphs with a given number of nodes), and connected components (maximally connected subgraphs with a given number of nodes).

Attribute-based criteria use the semantic properties contained in nodes and edges in order to attract nodes. Users can set a magnet to attract all the nodes (or edges) in which a certain property exists, or not only exists but is also equal to a certain value or is within a certain value range (if it is numerical).

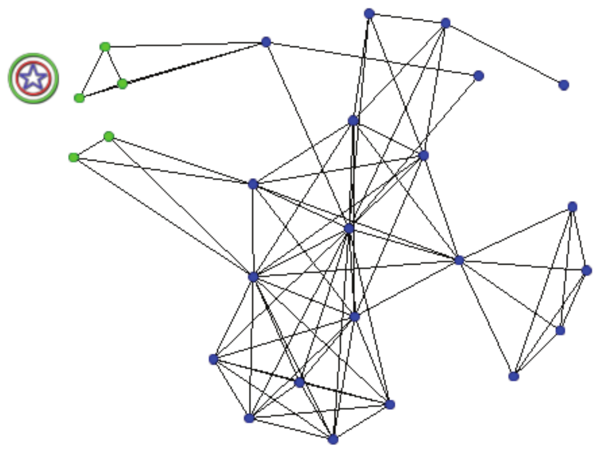

Figure 1. A magnet in action

With magnet-based criteria, one can set a magnet to attract all the nodes that another magnet also attracts, all the nodes that another magnet does not attract, all the nodes that no magnet attracts or all the nodes that are attracted by a combination of magnets. This allows for set-based operations on the graph visualization.

Within the physics metaphor, a magnet works simply, on each frame, by applying to all attracted nodes a force vector in its direction with the specified magnitude. Also, to keep the magnet from being overlapped by its attracted nodes and to keep the attracted nodes from staying all bundled together too close to each other, the magnet also exerts a repulsion force on each of the attracted nodes. This repulsion force is the same as with common nodes, working like a reverse gravity by being inversely proportional to the distance of the node to the magnet and proportional to the magnitude of the force of attraction, so that it is stronger with the nodes that are near the magnet and weaker with the ones that are progressively further away from it. It is also possible to assign a color to all the nodes that a magnet attracts.

3.3. Boundary Shapes

A boundary shape is simply a geometric shape (a circle in the prototype), which can be placed around a magnet and have the function of binding the nodes that such magnet attracts to the region that the shape delimits. At the same time that all attracted nodes are kept within the boundary shape, all other nodes are kept out, with the shape exerting a repulsion force similar to the one exerted on the other nodes by the nodes themselves (a reverse gravity force). Once a node finds itself inside a magnet's boundary shape, it cannot escape that area unless it is also attracted by another magnet that is placed outside such shape.

3.4. Magnet Hierarchy and Intersections

A magnet effectively creates sets of related nodes and ensures that they remain near a certain physical region. Occasionally it might be useful to refine this set of nodes into subsets. To allow for that, it is possible to define magnets that act only on the subset of the graph that is already attracted by another magnet. To do this, the user must simply create a magnet and define another one as its parent. It is interesting to note that child magnets might have child magnets of their own; creating thus a hierarchy of magnets that might be helpful for incremental exploration of a graph.

Occasionally it may happen that a node fulfils the criteria of two or more magnets. In such case, the user will be asked if the common nodes should be left free, with both magnets exerting their forces on them, or if they should be bound to any of the magnets. If they are bound to a magnet, they will be treated like any other of that magnet's nodes.

Nodes that are part of intersections are visually distinguished from regular nodes by having a different icon and by colored lines that link them to the magnets that attract them. Each line is in the same color as the nodes of the magnets that attract them.

4. Visualizing Co-Authorship Networks

In this section we show how the previously described visualization technique can aid in the exploration of social networks – more specifically, a co-authorship network.

The dataset here utilized was taken from all the works published in 2007 by the professors and students of the Informatics Institute of the Federal University of Rio Grande do Sul and their external collaborators. All publications in journals, conference proceedings and books are covered.

The graph built with this dataset contains 474 nodes and 1252 edges. Each node represents an author while each edge represents all the publications between two authors. As attributes, each node contains the author's name, degree (number of edges connected to it), total number of publications, number of publications in conference proceedings, number of publications in journals, number of publications in books and their category (whether they are faculty members, students or external collaborators). Each edge contains the id of their two nodes, the years of their publications, the types of the publications (journals, conference proceedings or book) and the number of common publications. The generated graph can be observed in Figure 2.

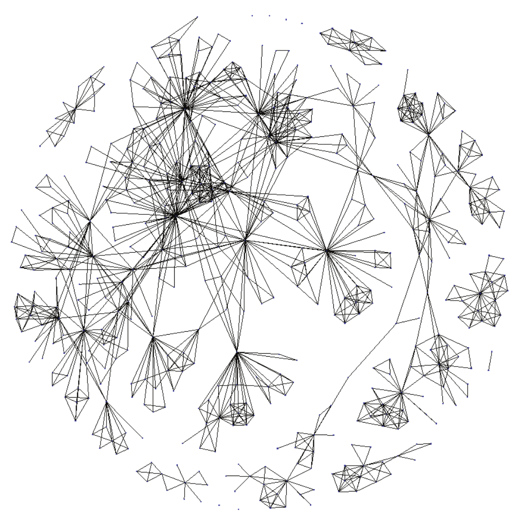

Figure 2. UFRGS 2007 co-authorship network

For this case study, we compiled a list of questions pertinent to the exploration and analysis of co-authorship networks. In the following sections, we show how they can be answered using the proposed visualization technique. Note that the answers provided for each question are just suggestions – using these tools, a user might find different ways to find results for the same question.

Since our prototype was used as is, without being adapted in any way for the specific case of co-authorship network visualization, some of these questions would require small adaptations of the prototype software to be answered with the proposed technique. Whenever that is the case, we point out exactly why the generic version of the technique cannot be used and which modifications are needed to better reach a solution.

4.1 Faculty member with the largest number of collaborators

The number of collaborators of any author is their node's degree on the graph. Therefore, to find out which faculty member has the largest number of collaborators we can insert in the scene a magnet that attracts the node with the highest possible degree amongst the nodes of faculty members. To do so, we first create a magnet that has a requirement to attract all nodes that have the attribute "category" with the value "faculty" and then create another magnet parented to the first one that has a requirement to attract the node with the highest possible degree. The result can be seen in Figure 3.

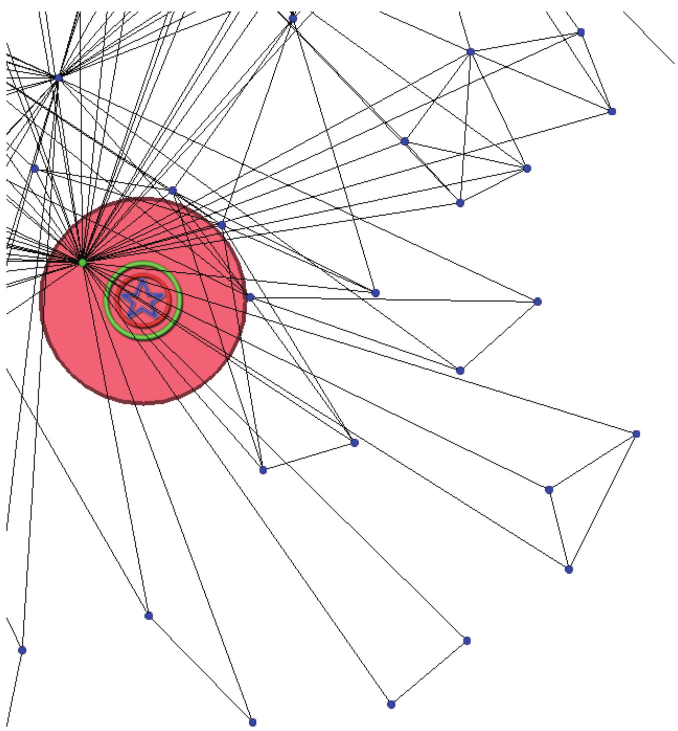

Figure 3. Faculty member (Ricardo Reis) with the most collaborators (51).

The current implementation lacks a criterion/requirement for the attraction of the node with the highest possible degree, so in the prototype one has to use a regular degree criterion and keep editing it until the largest is found. However, the highest possible degree criterion is a small modification that will be implemented soon.

4.2. Faculty member with the largest number of external collaborators

To discover who amongst all faculty members has the largest number of external collaborators, we could create a magnet that attracts all faculty members and another one that attracts all external collaborators. We could then create yet another magnet parented to one that attracts faculty members and set it to attract the node with the highest possible amount of edges to nodes that are attracted by the external collaborators magnet.

The prototype does not include yet a criterion/requirement for the attraction of nodes that have a certain number of edges (within a given range or the highest possible) to nodes that are attracted by a certain magnet. Using this same implementation, the answer to this question could be obtained by changing the dataset to include as attribute of a node the number of collaborators of each category.

4.3. Students that have worked with two given professors

To answer this question, we can first find out all the collaborators that two professors have in common and then see, amongst these, who are students. We can do this by creating a magnet that attracts all of the first professor's collaborators by using a requirement that targets all nodes that have path length of 1 to this professor. We then create a similar magnet for the second professor – with the difference that this one is parented to the first's magnet. This way we find all of the first's professor's collaborators that are also collaborators of the second, and have "student" as their "category" attribute.

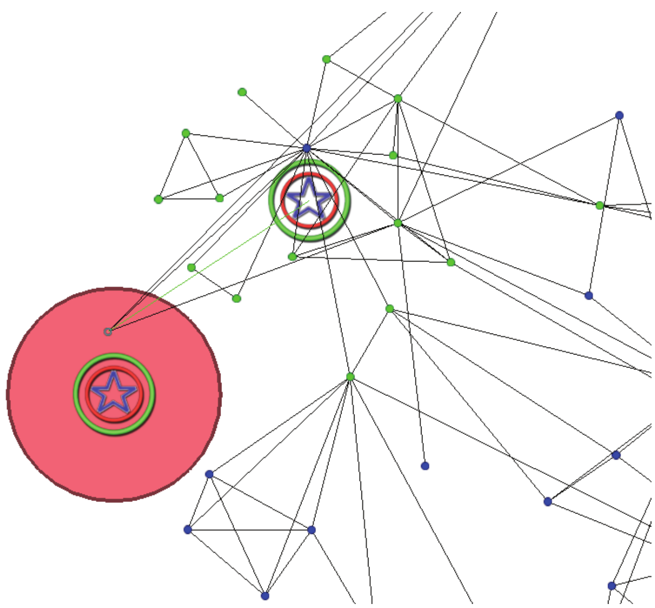

Figure 4. Common student collaborators of two professors.

In Figure 4, Luciana Nedel's collaborators are the light nodes attracted by the magnet inside the circle. The other magnet attracts nodes representing co-authors of Carla Freitas. All of Luciana´s collaborators who are students and also collaborate with Carla Freitas are attracted to the boundary shape since we set these nodes to be attached to Luciana´s magnet: only one student (Marta Villamil) is the answer for the question.

4.4. Student with the most publications in journals

We could find the answer to this question by creating a magnet with two requirements: one to attract all nodes with the “category” attribute being a student and the other to attract the node with the highest value for the attribute “journals.” This would result in the single node that represents the student with the most publications in journals.

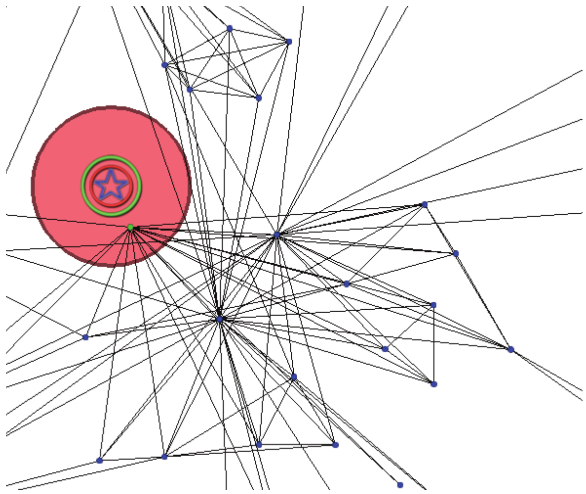

Figure 5. Student with the most publications in journals in 2007 (L. Agostini)

Since in the current implementation we do not have an attraction criterion/requirement for the node that has the highest amount of a certain attribute, we can answer this question by finding all the students who have at least one publication in journals and keep refining our search from there by then looking for students who have at least two publications, and so on. The answer to this question can be seen in Figure 5.

4.5. Researchers with publications in both conference proceedings and journals

This question can be answered by simply inserting a magnet with two requirements, one for all nodes that have the attribute “journals” higher than 0 and the other analogous, but for the attribute “proceedings.” The answer to this question can be seen in Figure 6 as the nodes inside the boundary shape.

5. Final comments

The goal of this paper was to show how a graph visualization application can aid in the study of social networks. For that, we have used a visualization technique [Spritzer and Freitas 2008] to investigate the co-authorship network of all the 2007 publications by the members of the Informatics Institute of UFRGS. We then showed how a few questions would be answered using the mentioned technique, following what was discussed in a previous work [Freitas et al. 2008].

Beyond the questions discussed in Section 4, we have investigated other questions like “Which groups did not collaborate with other researchers? (Figure 7) and “Which external collaborators have worked with more than one professor?” or “Is there any professor that has not collaborated with any student?”

In its current state the prototype does not support the two last questions. To be able to answer them, it would need a criterion/requirement for the attraction of nodes that have a certain number of edges (within a given range) to nodes attracted by another magnet. Although the technique would support that, the current prototype does not.

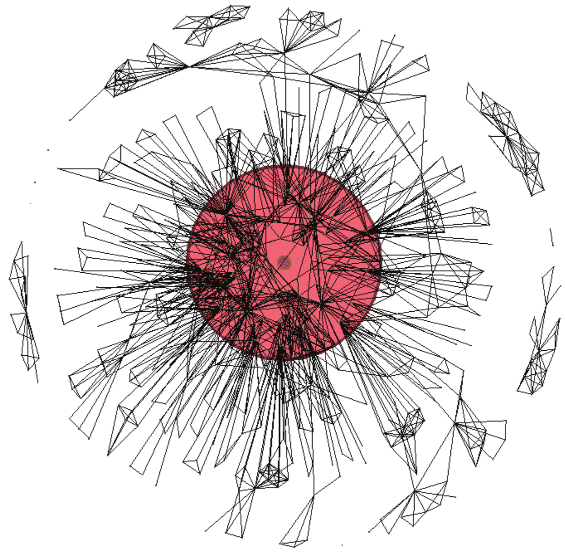

Figure 6. Researchers with publications in journals and proceedings.

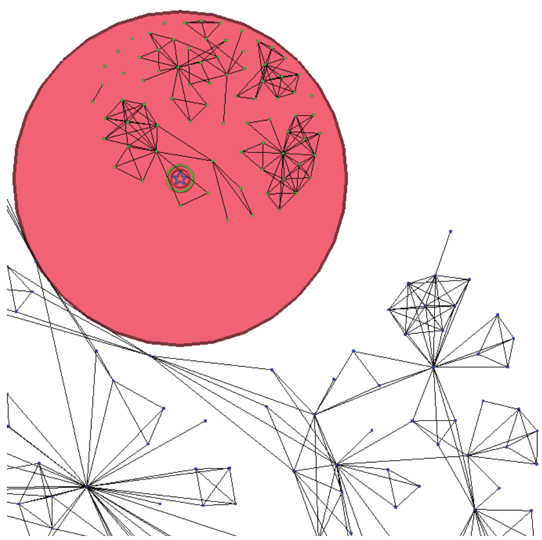

Figure 7. Groups of less than twenty authors that did not collaborate externally.

One interesting experiment would be to try and visualize the dataset representative of the social network made up by a more complete database of the same group but along many years, or of the national scientific production (such as in the Lattes platform). This would involve much larger datasets and would allow for a deeper study of the behavior of researchers based on their past work. Through visual and interactive queries, one might be eventually able to find out potentially interesting and useful behavioral patterns of successful researchers and evaluate one's production using more trustworthy and efficient methods.

6. Acknowledgements

We would like to thank Luciana Nedel and Renata Galante for their suggestions concerning the case study and our colleagues for testing our application. We would also like to thank Jean Daniel Fekete and Nathalie Henry for their helpful suggestions and for having provided the dataset used throughout the development of our technique.

6. References

Abello, J. and Korn, J. (2002) "Mgv: a System for Visualizing Massive Multidigraphs," IEEE Transactions on Visualization and Computer Graphics,vol. 8 (1), pp 21–38, 2002.

Barnes, J. and Hut, P. (1986). "A hierarchical O(N log N) Force-calculation Algorithm," Nature, v.324, n.6096, p.446–449, Dezembro1986.

Freitas, C.M.D.S, Nedel, L.P., Galante, R. et al. (2008) "Extração de Conhecimento e Análise Visual de Redes Sociais". Proceedings of SEMISH 2008.

Fruchterman, T.M.J. and Reingold, E.M. (1991) "Graph Drawing by Force–Directed Placement. Software - Practice & Experience", vol. 21, 1129–1164.

Heer, J. and Boyd, D. (2005) "Vizster: Visualizing Online Social Networks." In Proceedings of InfoVis 2005, IEEE Symposium on Information Visualization, Outubro 2005, 32-39.

Henry, N., Fekete, J.-D. and McGuffin, M. (2007a) "NodeTrix: Hybrid representation for analyzing social networks". IEEE Transactions on Visualization and Computer Graphics, 13 (Nov-Dec 2007), 1302-1309.

Henry, N., Goodell,H., Elmqvist, N. and Fekete, J.D. (2007b). "20 years of four HCI Conferences: A Visual Exploration". International Journal of Human-Computer Interaction, 23(3):239-285.

Herman, I., Melancon, G. and Marshall, M. S. (2000) "Graph visualization and navigation in information visualization: A survey," IEEE Transactions on Visualization and Computer Graphics, 6 (Jan-Feb 2000), 24-43.

Perer, A. and Shneiderman, B. (2008) "Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis", Proceedings of CHI 2008, ACM Press, 265-274.

Spritzer, A.S. e Freitas and C.M.D.S. (2008) "A Physics-based Approach for Interactive Manipulation of Graph Visualizations," Procedings of the Working Conference on Advanced Visual Interfaces, AVI 2008, 271-278.

Tunkelang, D. (1999) "A Numerical Optimization Approach to General Graph Drawing," PhD Thesis, School of Computer Science, Carnegie Mellon University.

Wasserman, S. and Faust, K. (1994) "Social network analysis: methods and applications", vol. 8 of Structural analysis in the social sciences. Cambridge Univ. Press, Cambridge.