Martins RM, Andery GF, Heberle H *et al.* Multidimensional projections for visual analysis of social networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(4): 791–810 July 2012. DOI 10.1007/s11390-012-1265-5

Multidimensional Projections for Visual Analysis of Social Networks

Rafael Messias Martins[1], Gabriel Faria Andery[2], Henry Heberle[1], Fernando Vieira Paulovich[1], Alneu de Andrade Lopes[1], Helio Pedrini[3], *Member*, *IEEE*, and Rosane Minghim[1], *Member*, *IEEE*

[1] *Institute of Mathematics and Computer Science, University of São Paulo, São Carlos 13566-590, Brazil*

[2] *CPM Braxis Capgemini, São Paulo 04543-900, Brazil*

[3] *Institute of Computing, University of Campinas, Campinas-SP 13083-852, Brazil*

E-mail: {rmartins, gfandery, henryhbrl}@gmail.com; {paulovic, alneu}@icmc.usp.br; helio@ic.unicamp.br; rminghim@icmc.usp.br

Received March 15, 2011; revised January 12, 2012.

Abstract Visual analysis of social networks is usually based on graph drawing algorithms and tools. However, social networks are a special kind of graph in the sense that interpretation of displayed relationships is heavily dependent on context. Context, in its turn, is given by attributes associated with graph elements, such as individual nodes, edges, and groups of edges, as well as by the nature of the connections between individuals. In most systems, attributes of individuals and communities are not taken into consideration during graph layout, except to derive weights for force-based placement strategies. This paper proposes a set of novel tools for displaying and exploring social networks based on attribute and connectivity mappings. These properties are employed to layout nodes on the plane via multidimensional projection techniques. For the attribute mapping, we show that node proximity in the layout corresponds to similarity in attribute, leading to easiness in locating similar groups of nodes. The projection based on connectivity yields an initial placement that forgoes force-based or graph analysis algorithm, reaching a meaningful layout in one pass. When a force algorithm is then applied to this initial mapping, the final layout presents better properties than conventional force-based approaches. Numerical evaluations show a number of advantages of pre-mapping points via projections. User evaluation demonstrates that these tools promote ease of manipulation as well as fast identification of concepts and associations which cannot be easily expressed by conventional graph visualization alone. In order to allow better space usage for complex networks, a graph mapping on the surface of a sphere is also implemented.

Keywords social network, visual exploration, multidimensional visualization

1 Introduction

The motivation for the study of social networks is varied, but the principle is common. There is a great amount of information from human relationships in social networking sites and databases that can be used for various purposes, such as to investigate preference patterns to support commerce and production sectors, to detect and investigate illegal activities, and to discover new forms of communication among individuals.

The most common way to represent social networks is using graphs, visually represented as diagrams containing vertices and edges. Nodes are actors and edges are their relationships. Communities are groups of individuals connected together, represented by connected sub-graphs. Various algorithms for graph display exist, and most of them are based on force-based strategy or on cluster detection algorithms. These approaches have the drawbacks of slow processing for large graphs and of sensitiveness to local minima or maxima.

Several tools for visual analysis of social networks use graph representation, as described in [1] and [2], which highlight relations between actors and groups of actors. However, data from social networking sites have attributes in the vertices and edges, which can be numerical, temporal or textual, and multiple relationships between the vertices. Some of these characteristics have received attention lately, but additional tools are needed to support exploration of behavior patterns affected simultaneously by the structure of the network (the connections), the properties of individuals and the properties of the communities. If the graph layout itself were capable of taking into account these attributes, data exploration would benefit from better node distribution regarding important data features.

In this paper, we propose and evaluate the concurrent use of two novel approaches that support data centered visual exploration of social networks. Taking the

Regular Paper

heterogeneous network as a starting point, in which vertices represent both individuals and communities, the goal is to promote easy visual location, in the layout, of groups of nodes that are related by their properties. The first approach places the network nodes on a plane by projecting them onto the visual plane according to similarities of their attributes, their connectivities, or both. To fit the projection requirements, a solid strategy is devised to translate the various possible types of attributes to numerical values. More than one view of the same graph can be generated at once to provide views of different contexts on the same data. In the second approach, the resulting views can be coordinated to crossly analyze attribute against structure in the social network, or even to associate properties of different networks with shared individuals. Additionally, graphs can be mapped onto a surface of a sphere, a view that has some advantages in space occupation and on added ability to interpret based on visual attribute mapping.

These approaches are implemented in a freely available system that offers these and other functionalities to ease exploratory analysis of social behavior. Results have shown that nodes are well placed regarding their properties, supporting focusing on regions of interest. Users can quickly associate individual attributes and similarities to their organization in communities, or associate content of their production to professional relationships. Besides being reflected visually on the layouts, some of these claims are supported by a user evaluation procedure that we carried out, whose results are also presented here.

This paper is organized as follows. Section 2 describes the work related to the visual analysis of social networks. The proposed approaches are detailed in Section 3. Section 4 presents the results in the form of case studies, user evaluation, and description of some important functionalities of the tool PEx- Graph. Conclusions are presented in Section 5.

2 Related Work

Several papers have been published describing visual exploration methods for social network data. The most intuitive and also the most common approach employs graphs to represent such data, since social networks have the intrinsic characteristic of being formed by nodes and edges, where the nodes represent network actors and the edges represent relationships between them.

Some tools for analyzing various fields using such elements are described by Huisman and Duijn[2]. Among them are: NetDraw, able to visualize large social networks; StoCNET, MultiNet, UCINet and Agna, which perform statistical analysis on social networks; Blanche and Condor, capable of simulating the network evolution. The paper also describes a library in *R* package for statistical analysis in social networks. However, despite the large number of existent tools, few have features to represent the attributes of the vertices, and those that have, merely change their visual properties, such as shape, color and size.

Other approaches have been developed to analyze specific areas. Heer and Boyd[1] presented a tool, called Vizster, which allows end users to identify patterns and have a more comprehensive view of the communities to which they belong. The positioning algorithm used is based on spring forces. The approach also allows users to modify the color of the vertices to reflect some attribute of the network actors, and to identify community structures from the edges.

MatrixExplorer[3], NodeTrix[4] and Tulip[5] are three interesting tools. The first one allows users to view the network as a set of nodes and edges in coordination with a matrix representation. The tool also creates a view of connected components, where each component is viewed as a compact rectangle whose size and color reflect the number of vertices contained in the component. The second tool represents the network as a set of nodes and edges, in which nodes are shown as adjacency matrices. For instance, each node can represent one community, the relationship between community members is represented in the adjacency matrix and inter-community relations are represented as edges of the graph. The visual properties of the vertices and edges can be modified to reflect attributes of the network. The third tool is an information visualization framework for the analysis and exhibition of relational data. The framework enables the development of algorithms, data models, interaction techniques, visual encodings, and domain-specific visualizations.

Namata *et al.*[6] also presented a tool, called DualNet, capable of generating coordinated representations. The tool treats the network as a set of sub-networks, each of which can be viewed and manipulated independently in different coordinated panels. The tool also allows the modification of visual properties of the vertices to reflect the attributes of the network.

A similar work to one of the approaches proposed in this paper is described by Shen *et al.*[7]. The tool developed by them, OntoVis, displays heterogeneous networks, that is, networks in which vertices represent more than one type of object. From an ontology graph, containing the different types of objects, users can construct a derived graph from the original graph by including only nodes whose types are selected in the ontology graph. The size of the vertices may reflect some measure, as node degree or centrality, and the color indicates the type of the vertex.

Perer and Shneiderman[8] described a system, called SocialAction, that uses attribute ranking and coordinated views to help users examine a number of social network analysis measures. Users can iterate through visualizations of measures, aggregate networks using link structure, find cohesive subgroups, and focus on communities of interest, as well as separate networks by viewing different link types individually, or find patterns across different link types using a matrix overview.

Aris and Shneiderman[9] developed the Substrate Designer, a tool for users to specify attributes for grouping nodes into non-overlapping regions, and attributes for placing them within regions. Users can specify the placement algorithm and decide on additional visual parameters, facilitating many network analysis tasks.

Li and Lin[10] proposed a mechanism for egocentric information abstraction in heterogeneous social networks. They extract a set of features from a given ego node based on linear combinations of its relations, and calculate statistical dependency measures between these features and the ego node. After filtering, they generate a condensed feature graph representation as the abstraction of the given ego node.

A few social network analysis tools have been concerned with displaying nodes according to attributes or similarities, a proposal similar to ours. One approach is described by Gloor *et al.*[11], which aims at visualizing social networks as a graph in which the positioning of the vertices is based on the similarity of the content of the messages exchanged by the actors. Velardi *et al.*[12] also presented an approach to grouping nodes based on the content of the messages exchanged by the actors of the network. Graphdice[13] employs various multidimensional visualizations in support to visual analysis of social network, including one type of projection. Here, we suggest the use of other types of projection that are meant to organize the display by similarity, which can be calculated from various types of attributes, including connectivity, and in this way contrasts and is complementary to their strategy. In our strategy, the dimensions themselves are transparent to the users at positioning, though driving the display.

Smith *et al.*[14] proposed an approach based on attributes to computing a degree of similarity between actors. For each attribute, a network can be generated in which the weight of the edges reflects the degree of similarity calculated for that attribute. Other approaches to representing graphs based on data contained in the vertices and edges are described by Pretorius and Wijk[15], and Archambault *et al.*[16]. Those approaches define a hierarchical structure based on the attributes to view the network.

The work of Wattenberg[17] uses a scatterplot, in which the axes are determined by the dimensions of a query summarization. All the actors who take the same values for both axes are grouped into a single node, and its size reflects the amount of grouped actors.

One of the approaches presented in this paper uses multidimensional projection to position the vertices of the network according to a similarity measure calculated using either a measure of adjacency or all the attributes of the actors simultaneously, which may include the relations between actors. Furthermore, this work also proposes to coordinate the multidimensional projection view with a heterogeneous network view. While the heterogeneous network view facilitates the process of identifying relationship patterns established by the network itself, the multidimensional projection view promotes the placement of the nodes according to a similarity criterion based on the attributes of the actors.

By combining both strategies, we believe it is possible to increase the power of analysis of social networks that have attributes. We make available a tool that is capable of exploring social networks using these new approaches and also general network interaction features.

In next sections, we detail these new presentation strategies, as well as further strategies, functionalities and examples.

3 Multidimensional Framework for Visual Exploration of Social Networks

The approaches proposed in this paper allow 1) the visualization of networks with different types of vertices (see Subsection 3.1); 2) the visualization of networks based on multidimensional projection of the vertices when they are described by an array of attributes — in this case, the proximity between the nodes is an indicative of the similarity between them (see Subsection 3.2); 3) the visualization of network graphs based on connectivity similarity, whose drawing strategy includes a first step that projects, using a similar strategy as in 2), the nodes that have similar connectivity close to each other (see Subsection 3.2); 4) simultaneous analysis of related networks and projections of their contents via coordination (see Subsection 3.5).

These approaches were implemented in a tool named PEx-graph, an extension of a previously existing open source tool for projection-based visualization, the Projection Explorer — PEx[18]. Various graph exploration features were also added, such as the possibility to change the visual attributes of vertices, the generation of egocentric networks and various others.

3.1 Heterogeneous Networks

A heterogeneous network is a graph in which the

vertices represent more than one type of object. With this graph, it is possible to represent, for instance, relational data, such as n-m database tables. Fig.1(a) presents a graph whose vertices represent authors and edges co-authorship. Fig.1(b) shows a heterogeneous network in which the circle-shaped vertices represent the authors, the square-shaped vertices represent the papers, and the edges represent the relationship between authors and papers.

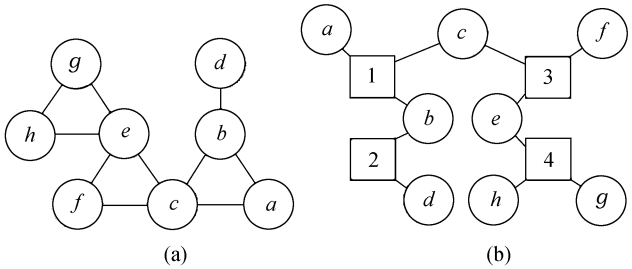

Fig.1. Co-authorship networks. (a) Only authors. (b) Authors and papers.

In this case, the user can choose to view the edges that connect vertices of the same type, as well as those edges that connect vertices of different types. When the graph represents individuals and communities (e.g., a paper can be seen as a community of authors), one can also explore individuals and their organization in a single display.

By using this graph along with typical interaction features in social network analysis, it is possible to do a much broader and more efficient exploration, according to the needs of the user. One of the most interesting features in this context is the ability to create the egocentric network of a vertex, which filters the view keeping visible only the vertex of interest and those which are connected to it. Furthermore, it is important to modify the visual properties of the vertices, such as shape, color and size, to represent the attributes of the data.

To illustrate this and the next approaches, we used a dataset of the Orkut social networking[19]. The dataset contains a table of individuals, consisting of an identifier, and the attributes "gender", "marital status", "birthday" and "age", a table of communities formed by an identifier and an attribute "name", and a table that associates the individuals to the communities to which they belong.

Fig.2 displays a heterogeneous network generated with the Orkut dataset, where the circle-shaped vertices represent communities, the square-shaped vertices represent male individuals, the single triangle-shaped vertex represents one female individual, and the edges connect individuals to their communities.

The positioning of the vertices was generated randomly and adjusted with a spring algorithm. For this example, the egocentric network of community "*Palmeiras desde Criancinha*" was initially generated (*Palmeiras* is a soccer team, and the community name in English is "Palmeiras Since Childhood"). This procedure allows us to identify that this community contains only one female individual. Later, the egocentric network of this individual was added to the previous network.

In our particular case, the network will include both individuals and communities in a single representation.

3.2 Multidimensional Projections

Each individual in a social network, typically represented as a vertex in a node-link diagram, can be described by an array of attributes, that is, as a vector in a high-dimensional space. This is known as the vector

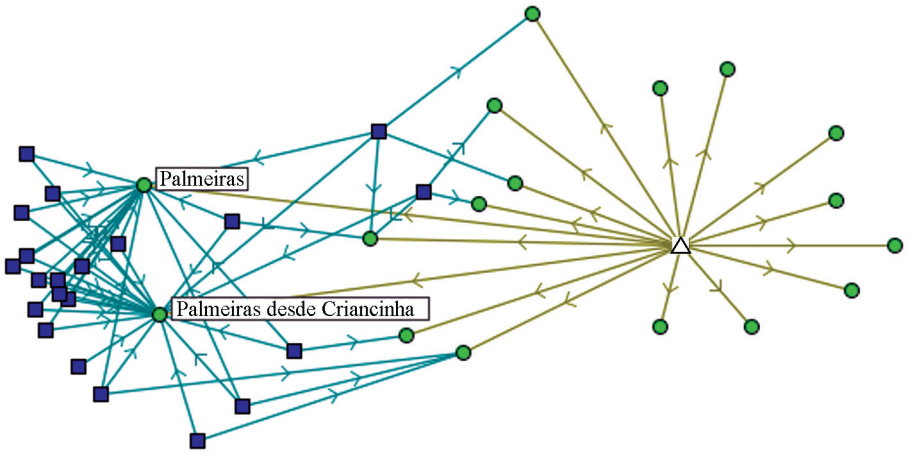

Fig.2. Heterogeneous network generated with the Orkut dataset, in which vertices represent individuals or communities and the edges connect individuals among themselves and to their communities.

space model, well known for text[20] and also widely employed for other unstructured data such as image, mostly in data mining and visualization applications. The multidimensional projection process reduces the number of dimensions of the dataset by mapping them into a low-dimensional space, generally two or three dimensions, while preserving as much as possible the distance relations between individuals. Here, we propose to employ the strategy for network data, particularly heterogeneous social network data.

The purpose of employing multidimensional projections is to place, on the visual layout, individuals that are highly related to one another close. The closeness criterion is usually given by a dissimilarity (or distance) relationship calculated on node attributes. When a network is the target object, nodes are highly related when they have similar attributes, or when they are similarly connected (that is, connected to the same nodes), depending on the perspective of the analyst. In either context, there are levels of similarity.

Attribute similarity will help locate profiles of subnets and is more appropriate for homogeneous networks, since different types of nodes have widely different attributes. Connectivity-based similarity, on the other hand, should support placing nodes with similar connections next to each other, which, in turn may help divide the graph into regions of highly connected nodes. That helps prevent arbitrary crossing of edges and improve display. Using multidimensional projection, it is possible to achieve this layout in one pass, that is, without resorting to graph analysis or force-based displays, which can be slow and more sensitive to small disturbances in the data.

In order to apply multidimensional projections to position nodes in a plane, each one must be defined by a series of numerical attributes. On top of that, a similarity relationship between individuals must be defined. The result is a distance matrix, from which a multidimensional projection can perform a similarity placement on a visual (2D or 3D) space.

In this work, we employed two ways to create the similarity relationship from node properties. One of them is based on the individuals' connectivities. The other is based on their attributes. The idea is to promote, on the layout, a favorable placement of the nodes, where neighborhoods on the display reflect similarity either in terms of their neighbors on the graph (connectivity) or in terms of sharing similar properties. Next, we describe how we code these two types of similarity as input to projection techniques.

3.3 Connectivity-Based Projection

The basic procedure that we adopted to produce a final (dis-)similarity matrix from node connectivity follows two steps. As mentioned before, a valuable form of node distribution via projection is by employing their connectivity as a means to calculate their similarity. The connectivity of the network, that codes relationships between nodes, can be seen as a node attribute. Thus, a distance matrix can also be calculated that reflects similarity between node connectivities.

The graph is built from the data by employing, as weights on the edges, the strength of their connections. Between the community and individual, weight is 1. Between individuals, it relates to the number of communities in common. Finally, between communities, it is the number of individuals in common. Naturally, this type of weight actually corresponds to similarity between the nodes. To reflect distance, it has to be reversed.

We have implemented two forms of calculation of an attribute distance matrix for connectivity. We call it relationship matrix, where:

1) the matrix contains the shortest path, on the graph, between each pair of nodes.

2) a modified adjacency matrix is calculated. In the adjacency matrix, each weight $a_{i,j}$ is replaced by $maxweight - a_{i,j}$, where $maxweight$ is the largest weight in the network.

This idea comes from the observation that the weight between nodes, once reversed, should in principle generate neighborhoods on the display, that is, similar connectivity should be reflected as proximity on the layout.

The example in Fig.3, in a smaller scale, illustrates the utility of the projection based on connectivity using the relationship matrix. This example is a hybrid network involving scientific papers and their authors. The connections are author-author (representing co-authorship), author-paper (representing authorship) and paper-paper (representing authors in common). In this case, a paper is a community of authors. The dataset is the bulk of published scientific production in the field of visualization in Brazil, for a period of 7 years (2003∼2010).

The network layout of Fig.3(a) confirms the previously observed fact that there are two large groups of researchers working on the theme: one in the southeast (top of the larger subgraph) and one at the south (bottom of the larger subgraph), with just one author cooperating between them in papers (in the middle part, the circle in the portion highlighted by color). Furthermore, the display is done only by placing the nodes, using IDMAP (interactive document map)[21], then connecting them with relationships edges. What the connectivity-based does is to prioritize the largest connected component, since there is a large difference in connectivity between that and the smaller connected

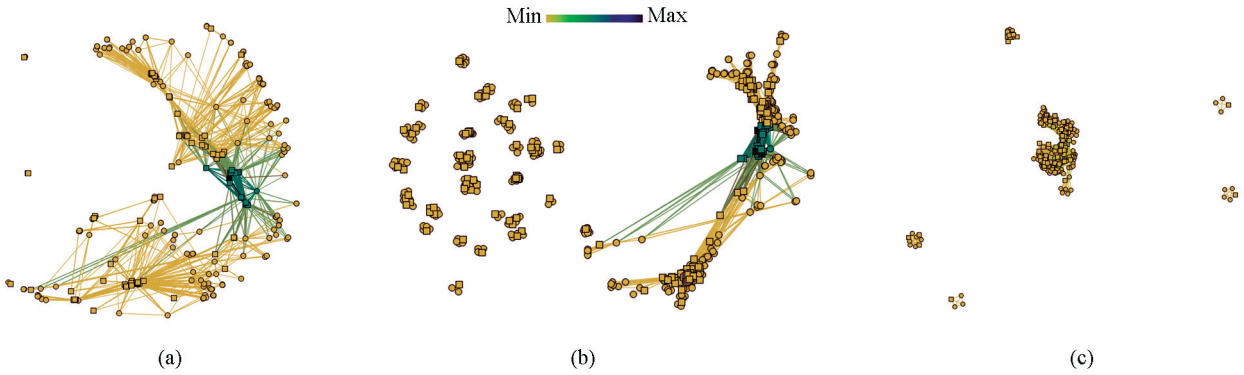

Fig.3. Relationship-based projection of a paper co-authorship network. Circles are authors and squares are papers. (a) Layout using IDMAP[21] projection only. (b) Application of very few steps of force-based layout on the graph in (a). (c) Conventional force-based layout.

components. The smaller components are concentrated on only a couple of spots.

Projecting nodes directly onto the layout prevents local minima from derailing force-based placements. When a force strategy is applied after projection of the distributed nodes, there is less likelihood of large groups of connected nodes crumbling together on the layout as it would happen with a conventional force-based placement with random initial layout. Section 4 will offer numerical evaluation of that property. Even a force-based placement, whose weight on the edges is given by connectivity calculation, cannot produce such separation effect. We have employed various graph layout software packages changing the decision on edge weights without being able to approach this type of node distribution. Fig.3(b) shows the result of force-based algorithm applied from the initial projected positions.

In the layout of the larger graph of Fig.3, the papers are placed around their authors, whereas the co-authors are placed in the same region. Close examination of this graph shows other interesting patterns. For instance, it is easy to observe that supervisors are laid out in the skeleton of the graph, and their supervisees in the surrounding region; a large amount of papers with the various common authors tend to clutter in the skeleton and push the authors out. These types of patterns are recurrent, so the approach will help users look for them when examining various exemplars of the same type of network.

Additional force-based movements of the nodes in Fig.3(a) reveals, to the left of the larger subgraph, a sizeable number of smaller graphs, representing individuals or smaller groups of researchers working isolated from other groups in the subject of visualization (see Fig.3(b)). The priority for the larger graph automatically caused by the projection allows subgraphs of smaller number of nodes to be quickly located and separated, so space is left to spread highly connected nodes. On the layout by force after projection (see Fig.3(b)), these smaller graphs are spread in one side of the viewframe and do not cause clutter at all. On the layout that is based on force placement only (see Fig.3(c)) these nodes are initially mixed with others and sometimes a large number of iterations is necessary to separate them, causing the larger graphs to clutter. As for interpretation, in this example, as in most graph-based bibliography analysis, there is a large connected component due to propagation of cooperation between researchers and the smaller graphs are smaller sets of papers, in isolated components, of research groups in other fields with explorations in that particular area, or of newly formed research groups. Both strategies for relationship matrices mentioned above lead to similar results of node distribution. We present further examples and numerical results in Section 4.

3.4 Attribute-Based Projection

The basic procedure that we adopted to produce a final (dis-)similarity matrix from node attributes follows two steps:

1) create a distance matrix for each attribute of the dataset. Each of these individual distance matrices contains a dissimilarity value, for each pair of vertices and for one of their attributes.

2) combine individual attribute distance matrix into one distance matrix representing the dissimilarity between every pair of nodes.

Step 1 of this procedure requires every attribute to be numerical. Next, we give an example of how this can be done for a number of sample attributes. In Step 2 of the attribute distance matrix creation, individual matrices are normalized between 0 and 1, and then summed up.

In order to create the individual attribute distance matrices shown in the example, it is necessary to

translate all attributes into numerical values.

The translation of node attributes to numerical values occurs as follows: 1) for string attributes, the Levenshtein distance[22] is used; 2) for numerical attributes, the difference between them is calculated; 3) for dates, the difference in milliseconds is calculated; 4) for textual entries associated with nodes, we use conventional word count strategies[23] for larger text entries or Normalized Compression Distance[24] for smaller text.

Table 1 shows an attribute matrix, in which each column is a data item representing a node of the network, and each line is an attribute.

Table 1. Data Matrix

	Node 1	Node 2	Node 3
Attribute 1	23	18	32
Attribute 2	2	3	6
Attribute 3	0	0	1
Attribute 4	Male	Female	Female

Table 2 shows the individual distance matrices calculated from each attribute of the dataset. For the three first attributes, translation was done employing rule 2) above, and for attribute 4 rule 1) above was used.

The distance matrices are summed up in Step 2.

As user's choice, the connectivity attribute matrix (see Subsection 3.3) may or may not be summed to the other attribute distance matrices, that is, it may be used isolated or in addition to the other attributes to contribute to node placement in the projection.

Using similarity matrices calculated in such form, any multidimensional projection technique can be used to layout nodes on the plane. However, some are more adequate in certain cases. For instance, for textual entries (e.g., when nodes are documents in co-authorship networks), LSP (least square projection)[25] has been shown to produce good grouping of highly related individuals for connectivity-based data.

However, for most cases we need a projection that separates groups but spreads them around the available area, decreasing cluttering. Various projection techniques[25-27] have been developed lately that aim at improving point placement regarding similarity relationships, while performing in real time. They are preferable to classical techniques, such as PCA (principal component analysis)[28], particularly due to the high number of attributes.

Fig.4 illustrates the projection based on the similarity matrix calculated with the summed attribute matrix of the Orkut dataset. The projection technique is called IDMAP[29]. The two images represent the same projection, but each one is colored according to one attribute. In Fig.4(a), the color represents gender: blue for male, red for female. In Fig.4(b), the color represents marital status: red for married, green for single, yellow for committed, blue for open marriage, cyan for open relationship. The projection by attribute has clearly grouped individuals by their attribute profiles, which may help investigate connections, when edges are drawn according to shared community preferences.

Attribute-based projection, such as the one proposed here, supports the analysis of social networks as well as other data. However, organizing node layout in a network by attribute similarity does facilitate finding relationships (edges) that may be reflected by such attributes, as shown in some examples of Section 4.

One point must be made clear regarding attribute mapping. Regardless of the normalization or standardization algorithm employed for attribute transformation needed for the purpose of projections, it is a fact that the large difference in variation of some attributes cannot always be smoothed. Although the differentiation procedure explained above helps soften influence of attributes with smaller sets of valid values, binary or categorical attributes will influence the layout differently than continuous or largely varying attributes. That property may actually help interpret the layout based on that particularly influencing attribute. For instance, if one has an attribute such as category and uses it to map data, that can clearly support the segregation of nodes on the layout based on that attribute. This helps find both other trends related to category and find exceptions, which would be the individuals in a category that do not fall close to the others in the same category. One example of that is given in Section 4.

To relieve the influence of binary or categorical attributes, it is possible to work with application-based attribute weighing or even to remove exceedingly influential attributes and re-map the data for analysis without influence of that particular attribute. In a preprocessing step, the system allows the user to choose what attributes should be kept or removed from their mappings.

Another possibility is to map based on connectivity,

Table 2. Distance Matrices for Each Attribute

	Attribute 1			Attribute 2			Attribute 3			Attribute 4		
	Node 1	Node 2	Node 3	Node 1	Node 2	Node 3	Node 1	Node 2	Node 3	Node 1	Node 2	Node 3
Node 1	0	5	9	0	1	4	0	0	1	0	2	2
Node 2	5	0	14	1	0	3	0	0	1	2	0	0
Node 3	9	14	0	4	3	0	1	1	0	2	0	0

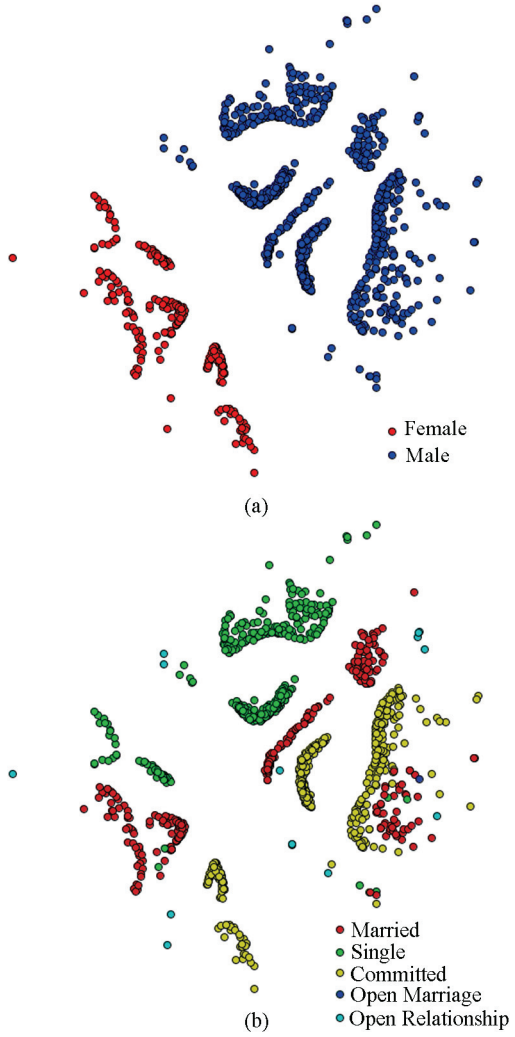

Fig.4. Multidimensional projection of the summed attribute matrix of the Orkut dataset. Color represents the attributes. (a) Gender. (b) Marital status.

using a target attribute as weight on the graph edge. Fig.5 exemplifies the projection from connectivity values of the Orkut dataset, with connectivity indicating whether two individuals belong to any common community. Weight is the number of communities in common. The projection technique is called IDMAP[21]. In this case, the vertices represent only individuals.

In the dataset, there are individuals in three different communities related to Peugeot cars. Vertices are colored according to the number of communities related to Peugeot to which the individuals belong. In this case, individuals who are not member of any of these communities appear in white.

As it is possible to see, the vertices representing those who belong to communities related to Peugeot are projected onto a different region than those who do not. The result allows the exploration and identification

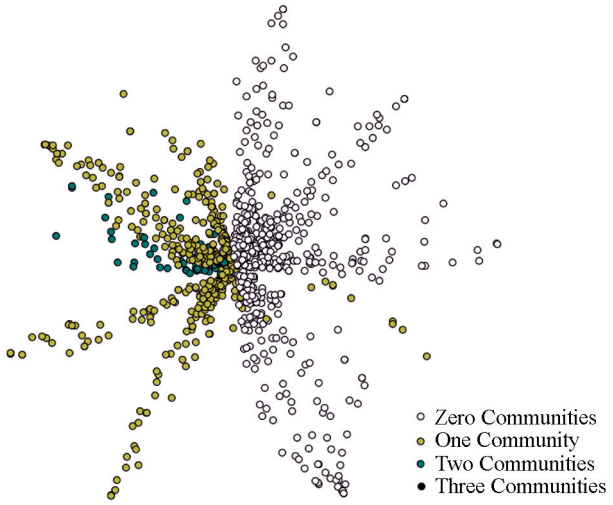

Fig.5. Multidimensional projection of individuals using the relationship matrix of the Orkut dataset, where color indicates to how many communities related to Peugeot individuals belong.

of other communities to which these groups of individuals belong and will prevent excessive crossing when the edges are shown, since all individuals that share the same communities also share the same region on the map.

3.5 Identity and Relational Coordinations

The third proposed approach uses an identity coordination mapping between different views created using the previous approaches. The identity coordination technique creates a link between the same vertices on different visual representations.

When a vertex or a set of vertices is selected in one visual representation, the same vertices are highlighted in the other representations. Thus, it is possible to identify groups of individuals by both object similarity (considering their attribute-value tables) and the connections of the network (considering the relationships between objects).

Fig.6 shows a coordination between two views

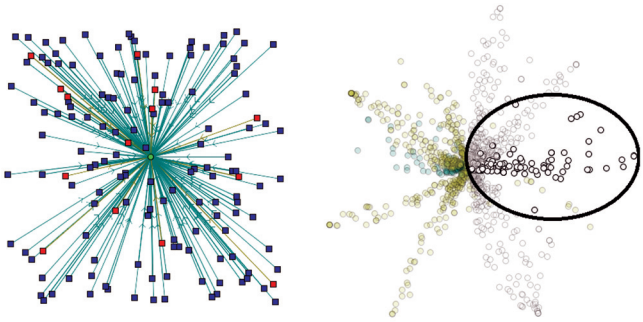

Fig.6. Coordination between two views created using the described approaches, where the members of the community "I hate Fiat!" were highlighted.

created using the previous approaches. From the heterogeneous network view on the left, it was created the egocentric network of the community "I hate Fiat !!", where the circle-shaped vertex indicates the community and the square-shaped vertices represent individuals.

In the multidimensional projection view on the right, the color is the same as in Fig.5, that is, it indicates to how many communities related to Peugeot the individuals belong. In this case, selecting the members of the "I hate Fiat!" community in the heterogeneous network view, they were highlighted in the multidimensional projection view, allowing users to check that none of them belong to any community related to Peugeot.

The relational coordination technique creates a link between vertices on different visual representations based on relational information. For instance, it is possible to create two visual representations, one of members and the other of communities, and then associate the members to the communities to which they belong.

4 Results

This section describes the results obtained by applying our methodology, including case studies, numerical results, functionalities of the tool and implementation issues, and a user study that was carried out.

4.1 Case Studies

The following subsections present two case studies for social networks and, associated with them, the various analysis capabilities of the techniques are presented here.

4.1.1 Netlog

For this case study, data were collected from members of two groups of the Netlog social networking[29]. The groups are "arsenal_fanz" and "Manchester_United_til_I_die". The attributes of the members are "sex", "country", "age" and "birthday". The groups of each member were also collected.

Fig.7 shows a heterogeneous network of the Netlog dataset. The groups are represented as green circle-shaped vertices, the female users as red square-shaped vertices and the male users as blue square-shaped vertices. The initial placement was random and then repositioned with a force-based placement algorithm[30].

The two main groups ("Arsenal_fanz" and "Manchester_United_til_I_die") in Fig.7 are marked with labels that identify them. By passing the mouse over the groups, a label displaying the group name appears. Various investigations are favored by this approach, such as common members of both communities, members exclusive to one of them, and other communities these members belong to.

Fig.8(a) displays a subset of the groups at the center of Fig.7, some of them marked with a label. Some of the groups common to the members of both are: "ITALY", "Salva L'Amore", "FRIENDS 4 U", "Linkin_Park" and "Al Pacino". Fig.8(b) displays the groups of which the members of "Manchester_United_til_I_die" are part. In addition to various groups directly related to the Manchester United football team, there are also two communities dedicated to the player Cristiano Ronaldo and others as "Top Gear" and "Nirvana". The groups related to the members of the "Arsenal Fanz" are shown in Fig.8(c).

Just as in the previous case, there are several groups related to the Arsenal football team, others related to football in general ("ACMilan" and "World Football") and others as "Final Fantasy".

Combining heterogeneous representation with multiple view coordinations, users can quickly switch between group interpretation — in terms of groups of members and of groups of similar communities — and interpretation of individual groups. The fact that the nodes are displayed based on similarity of connectivity allows for that type of interpretation. Although a conventional force-based view allows similarly connected neighborhoods to be reconstructed, such type of view would not be possible, since the large connection between many nodes would cause the graph to be cluttered by balanced forces between similarly weighted edges.

Fig.9 shows a multidimensional projection of the users considering both their attributes and their relations, which reflect the number of common groups between members. In Fig.9(a), the members are colored by country and, in Fig.9(b), the members are colored by age. Members on the left are male, and on the right, female.

It can be noted that there are many more male members than female. In addition, there are many members of the United Kingdom, as might be expected, since the groups are dedicated to football teams of that country, as well as members of Iran and Saudi Arabia. It is also possible to observe that most of the members are young.

Regarding the edges, a slider can control the display by weight; in this case, the weight reflects the number of common groups between members. Fig.10 shows the relations between the members with at least 10 groups in common. It is easy to see that the majority of the members who fit this case are from Iran and Saudi Arabia.

Analyzing other sets of users from other countries in more detail, including the United Kingdom, it is possible to see that the number of groups in common does not exceed 4.

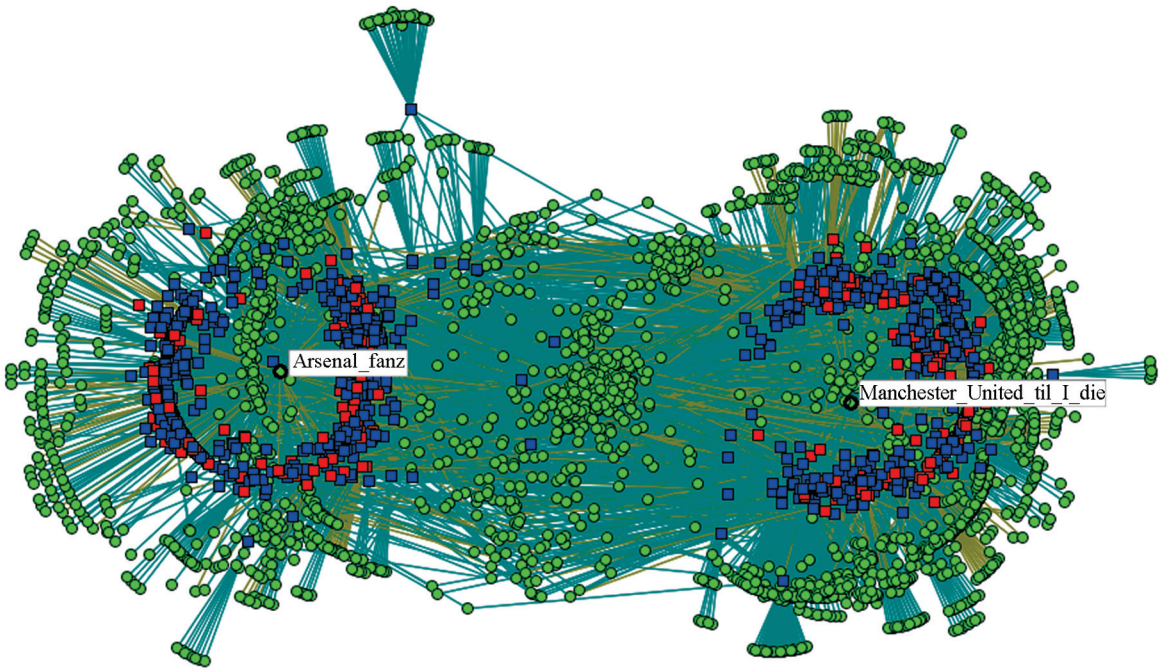

Fig.7. Heterogeneous network of the Netlog dataset.

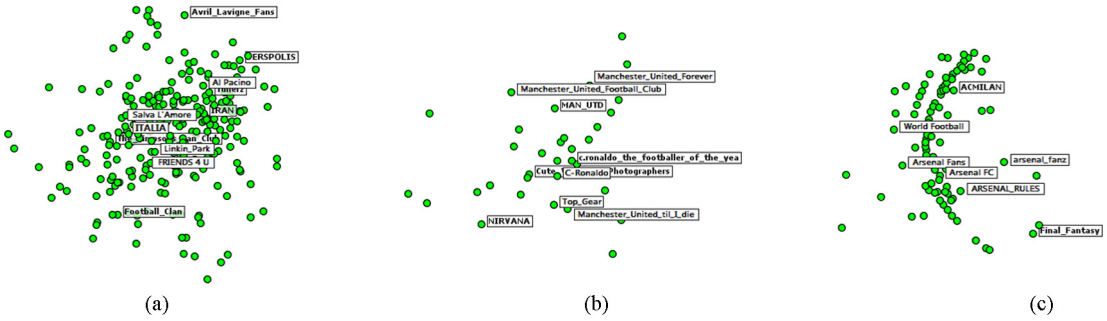

Fig.8. Subsets of the heterogeneous network. (a) Other communities signed by members of both "Arsenal_fanz" and "Manchester_United_til_I_die". (b) Other communities signed by members of "Manchester_United_til_I_die". (c) Other communities signed by members of "Arsenal_fanz".

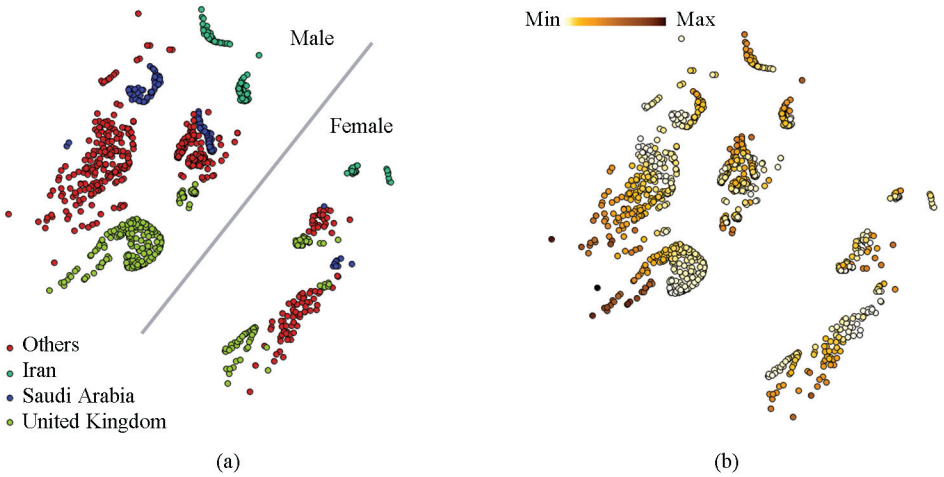

Fig.9. Multidimensional projection of the Netlog dataset users. (a) Colored by country. (b) Colored by age.

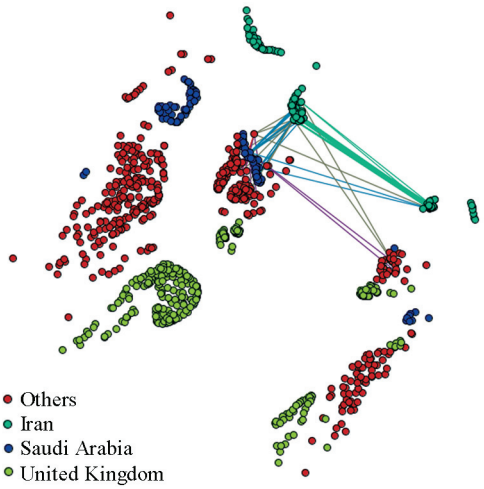

Fig.10. Multidimensional projection of the Netlog dataset users displaying the edges between members with at least 10 groups in common.

From analysis based on attribute, it is possible to explore even reasonably large networks trying to find associations between those attributes and the way that individuals connect. For example, attributes such as gender tend to separate clearly the individuals allowing to find other trends in attribute values and in connection values that may be related to that.

Since string similarity is available, one type of segregation clearly possible is to separate groups of members by tags they share and associate that with both their connections and their other attributes. This can be very useful for various applications such as tag-based recommendation[31]. An important point to note is that, if the user wants to remove the trend of a particular attribute from his or her analysis, all that needs to be done is to remove that particular attribute and the segregation caused by it will disappear.

Using the identity coordination feature, when selecting the users from Iran in the multidimensional projection view, they were highlighted in the heterogeneous network view, allowing to identify that most of them are part of the "Manchester United Til I Die" group, as shown in Fig.11.

4.1.2 Orkut

Using another subset of the Orkut dataset presented in Subsections 3.1, 3.2 and 3.5, we explored the profile of the members belonging to the communities that had the themes Peugeot and Volkswagen in the title.

Fig.12 shows a heterogeneous network projected using the adjacency matrix of the dataset. The relationships connect members who belong to at least 10 communities in common, communities which have at least 10 members in common, and members to the communities to which they belong.

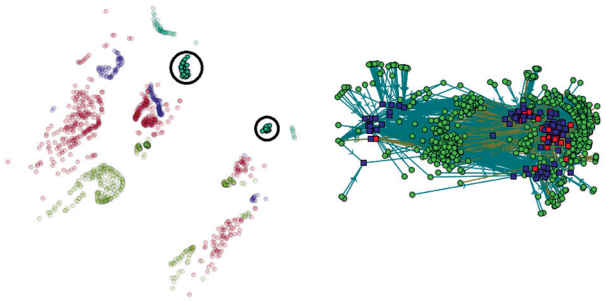

Fig.11. Coordination between the heterogeneous network view and the multidimensional projection view. The users from Iran were selected in the multidimensional projection view and they were highlighted in the heterogeneous network view.

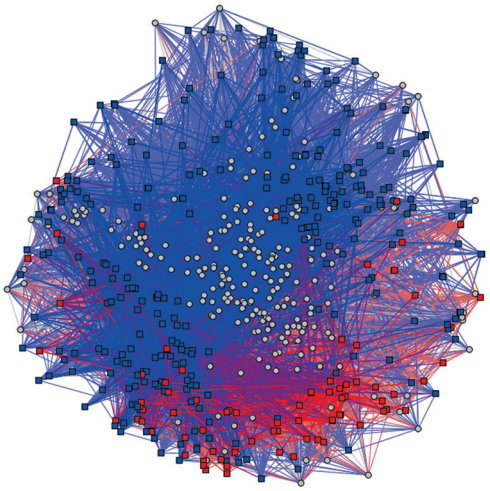

Fig.12. Heterogeneous network of the Orkut dataset.

The groups are represented as gray circle-shaped vertices, the female users as red square-shaped vertices and the male users as blue square-shaped vertices. It is possible to see that most of the female members are grouped at the bottom, along with the communities to which they belong.

Fig.13(a) shows the egocentric networks of the three Volkswagen communities, and Fig.13(b) shows the egocentric networks of the two Peugeot communities. We can observe that the communities related to Volkswagen have only male members, while the communities related to Peugeot have both male and female members.

Furthermore, the members of the communities related to Peugeot also belong to more romantic communities, such as "Carpe Diem", "Nothing happens by chance", "I love music" and "Lost", while the members of the communities related to Volkswagen also belong to more material communities, such as "Barbecue", "We want Coke 20 liters" and "The Simpsons".

Fig.14 shows a projection of the summed attribute matrix of the members. The color reflects their marital status.

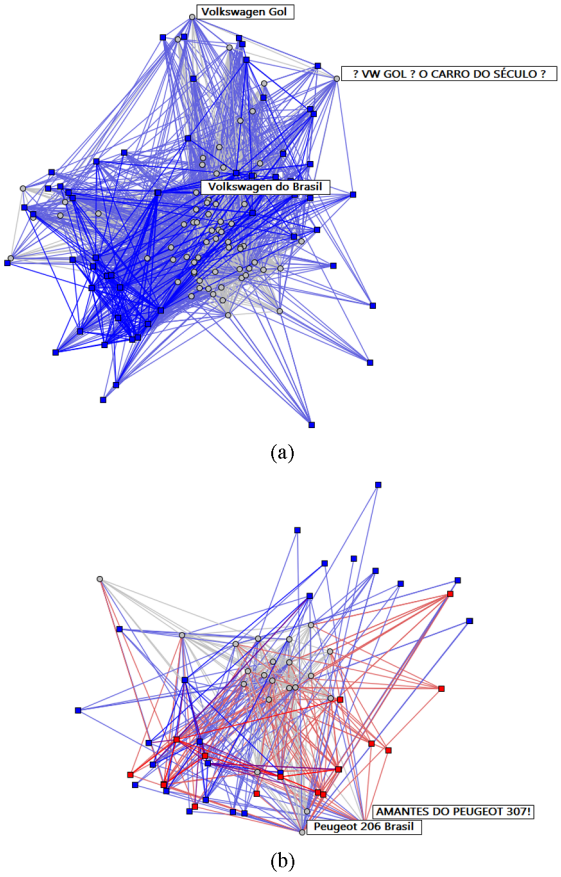

Fig.13. Egocentric networks. (a) Of the three communities for the theme Volkswagen. (b) Of the two communities for the theme Peugeot.

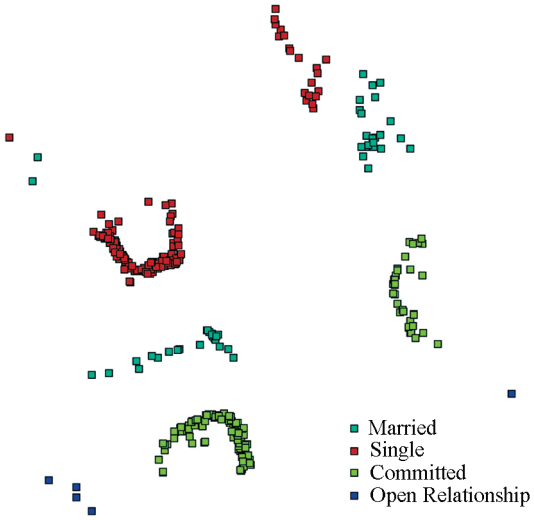

Fig.14. Multidimensional projection of the summed attribute matrix of the members for the Orkut dataset.

Figs. 15 and 16 show a relational coordination between this projection (on the left) and a projection of the relationships matrix of the communities (on the right). Thus, when we select a group of communities on the right projection, only their members remain visible on the left one.

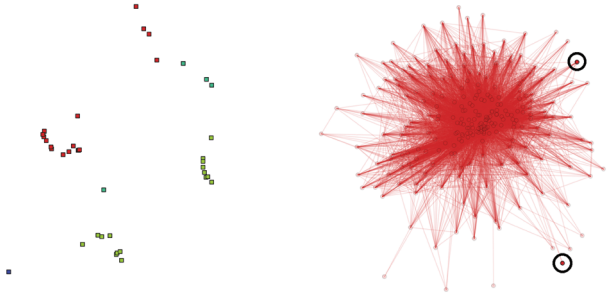

Fig.15. Relational coordination between the projection of the members and the projection of the communities. The communities related to Peugeot are selected on the right projection, and only their members remain visible on the left one.

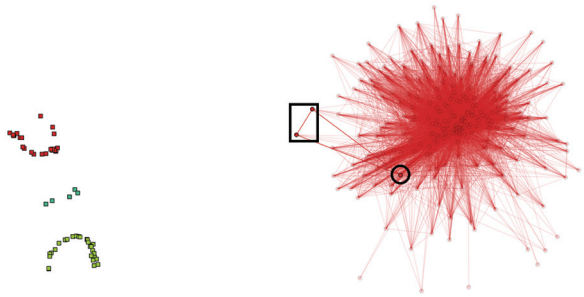

Fig.16. Relational coordination between the projection of the members and the projection of the communities. The communities related to Volkswagen are selected on the right projection, and only their members remain visible on the left one.

In Fig.15, the communities related to Peugeot are selected and, in Fig.16, the communities related to Volkswagen are selected. As we would expect, members of the communities related to Peugeot appear more scattered, while members of the communities related to Peugeot appear closer. This could give an indication for Volkswagen to target a new branch of customers.

Such mappings, that allow fast location of trend in community participation, can benefit many analysis tasks, such as product trend and location of novel behavior patterns for individuals or group of individuals.

4.1.3 Authorship Networks

The analysis of networks representing interactions among researchers and their research subjects is an application useful for several purposes, such as understanding the evolution of a subject area and evaluating impact, influence or performance from a certain point of view. This example is meant to illustrate the nature

of displays and types of analysis provided by connectivity based projections.

The dataset, recovered from [32], is the collection of all references of two major journals in area of computer graphics, the IEEE Transaction on Computer Graphics and Applications (TVCG) and Computer Graphics Forum (CGF). The full graph comprised 2 471 papers and 3 841 authors. The modified adjacency matrix was the criterion to perform the projection, via IDMAP, based on connectivity. Fig.17 shows various resulted graph views.

Fig.17 shows the first presentation given by the projection. The graph in Figs. 17(a) to 17(d) was built using the cosine distance metrics over the modified adjacency matrix. What happens in this case is that the distance compensates the size of the graph, highlighting the larger connected subgraph. The small point on the left in Fig.17(a) is actually the projection of all the other subgraphs. In this case, the graph nodes were colored by degree, highlighting, in colors away from red towards blue, the nodes with larger degree. If one looks at the projection behind the display, shown in Fig.17(b), it becomes clear that highly connected authors and papers are groups, even when they share connections with other individuals or groups, promoting useful local explorations. Also, that picture shows the coloring by a centrality measurement (betweenness), highlighting nodes (in colors green and blue) that are connectors between subgraphs across the whole graph.

In order to access the structures and sizes of the smaller subgraphs, it suffices to apply a force-based algorithm over this view. That was done in Fig.17(c). It can be seen that the relations change. The other graphs move toward occupying the larger portion of the layout. That is characteristic of force-based algorithms, though in this case a simplified version was used, moving nodes in small increments based on the distance between them.

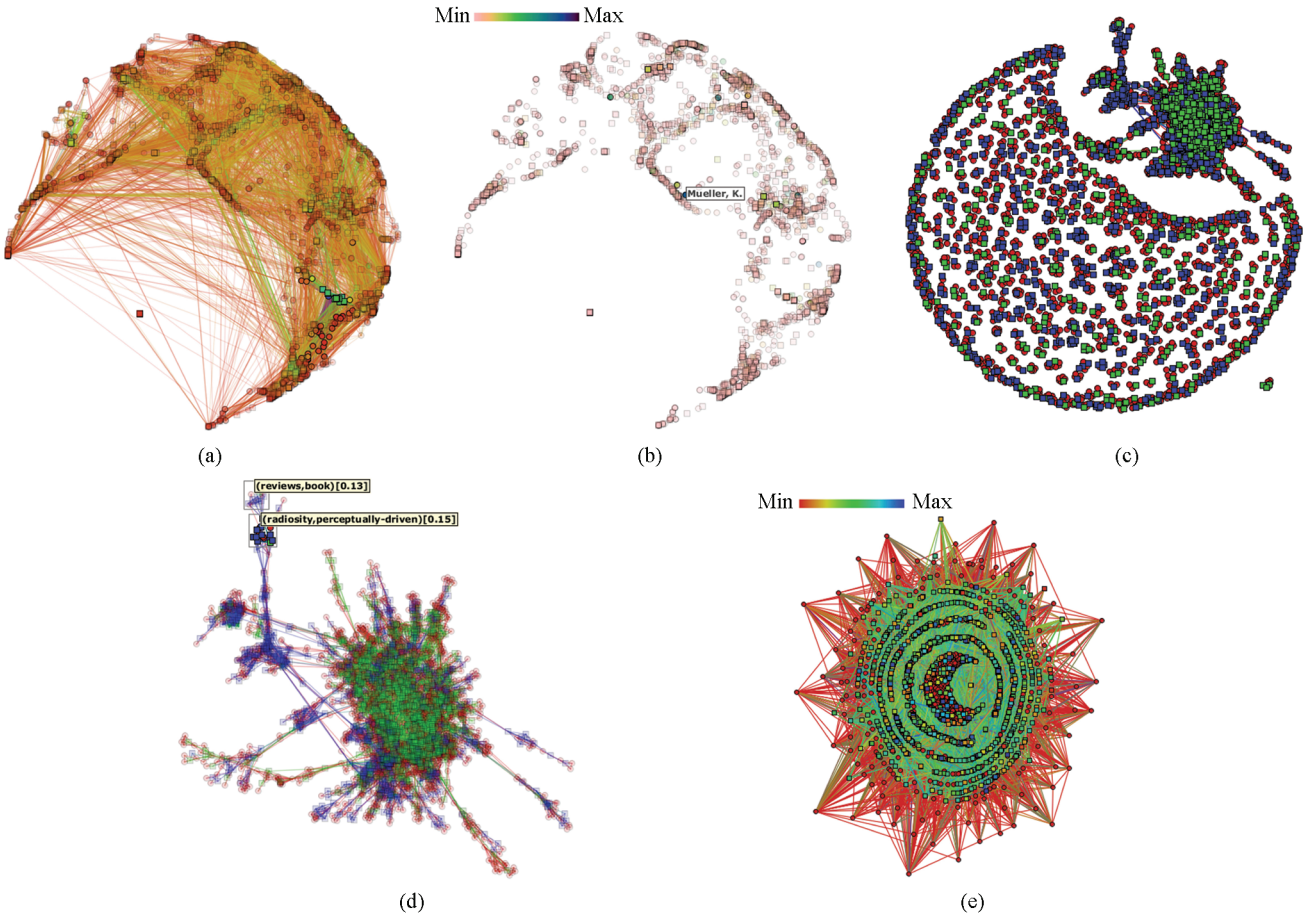

Fig.17. Paper-author network for all papers from TVCG and CGF journals composed of 6 312 nodes (2 471 articles). (a) Large connected graph corresponds to 3 692 nodes, where color represents the graph degree. (b) Result of projection, where color means betweenness. Transparency helps detect density. (c) Result after applying force to the graph in (a), where color represents journal. (d) Result of separating the larger graph from graph in (c), with topics of the top smaller subgraphs. (e) Display of the same graph using Euclidean distance, where peripheral nodes are the most connected nodes and color is year of publication.

Force-based movement of nodes also offers another perspective of the larger graph of Fig.17(a). That can be observed in Fig.17(d), which would be approximately the same result as applying a basic force-based graph layout. In that picture, we also employ another feature of our tool, the topic detection strategy. It is calculated from the words that co-occur most frequently in that set of nodes. It is possible to observe, on the top of Fig.17(d), two subgraphs related to book reviews and perception-based radiosity.

An interesting feature of our approach is seen in Fig.17(e). It illustrates the result of employing Euclidean distance to the same connectivity data as Figs. 17(a)∼17(d). By just projecting (via IDMAP) and connecting the edges, we get a star-like display of the graph, with layered nodes, that still highlight the groups of papers and authors and places the most connected nodes (in this case, naturally authors) in the outer layers. Such a star configuration is usually very slow to calculate using any other method. This layout lends itself to still another set of possible observations starting by the most connected individuals at the borders. It does not immediately shows the aspect of a larger subgraph as the previous layout, but that can be obtained by applying force and generating another display with spread out subgraphs, such as the one in Fig.17(c).

Next, we present evidence of point positioning by projection based on connectivity distributing the nodes on the display in a favorable way.

4.2 Analysis of Point Positioning

For the following tests, we employed three datasets of co-authorship networks. One of them (*eurovis*) includes all papers and authors of the proceedings of the EuroVis Symposium (and its precursor VisSym) from 1990 to 2010. The second dataset (*vis*) includes all papers and authors of proceedings of the conferences IEEE Visualization and IEEE InfoVis from 2001 to 2010. The third dataset (*agric*) includes papers and authors of a company that does research in agriculture from 2009 to 2011. All were fed into the system from a bibtex file of the papers.

Table 3 summarizes the sizes of the graphs formed. The last column is the percentage of nodes that belong to the larger connected component.

From the connectivity matrix calculated as defined in Subsection 3.3, initial tests lead to the decision of using cosine-based dissimilarity, since it was more successful in capturing the notion of two subjects being similarly connected.

Table 3. Testing Datasets

Dataset	Nodes	Edges	% Largest Component
Eurovis	1 285	4 838	59.1
Vis	4 435	21 175	72.2
Agric	6 178	30 965	73.2

Then, we generated the layout of the graphs, and, from that, extracted a number of measurements in order to establish what layout was favorable for disposing all the points in the available visual space, concerning the optimization of three criteria: distribution based on connectivity, number of line crossings, and capability of reconstructing clustered neighborhoods on the final visual space. Since the layout is dependant on the size of the visual space and number of iterations of the force-based algorithm, line crossings were counted with the same window size on the computer screen and the results averaged after 5 runs for each dataset and for each layout.

All measurements were taken from the force-based views, that is, the two most similar views were selected for analysis. One of them is the conventional force-based placement (force) and the other is the projection-based placement followed by minimal force displacement (proj-force), to confirm the improvement in positioning promoted by initial projection-based mapping. Two of the measurements were plotted and can be seen in Figs. 18 and 19.

One of the measurements executed is neighborhood preservation (*npres*) plot. Based on the attribute matrix for connectivity given in Subsection 3.3, the neighborhood preservation measures the percentage of neighbors on the layout that are the same neighbors as given by the connectivity matrix. The neighborhood preservation plots average for all nodes on the layout. Clearly, reproducing neighborhood in graph layout is not an easy task since no ideal placement can be met for highly connected graphs. But the plots in Figs. 18(d), 19(b) and 19(d) show that the layouts preceded by projections increase the possibility of points with similar connectivity being in the neighborhood of each other for all datasets.

Figs. 18 and 19 also show the advantage of projecting points before force placement when looking at the other measurement, the neighborhood hit (*nhit*). For calculating that precision measurement, first we cluster the points using a distance-based clustering method (k-means). Then, we count, for each node, the number of neighbors that fall into the same group of that note. The plot displays the average for all nodes. The colors in Fig.18 correspond to k-means clusters of the dataset. For all datasets, neighborhood based on groups of similarly connected nodes is improved when projection is performed prior to point positioning.

The top values for both *nhit* and *npres* are not large. The largest group reconstruction is 0.83 and neighborhood preservation is, at best, 0.425. The main

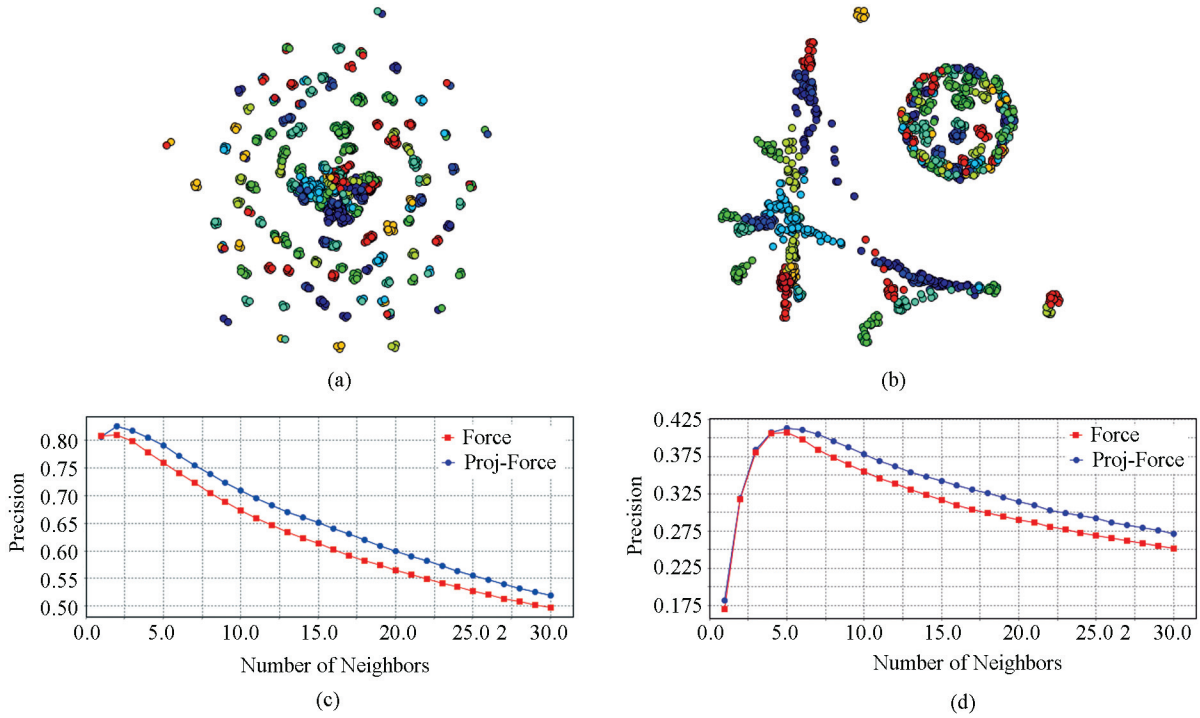

Fig.18. Point placements for eurovis dataset and their corresponding precision plots. (a) Force-based placement of nodes from random positions (force). (b) Projection followed by force-based placement (proj-force). (c) and (d) Neighborhood hit plot and neighborhood preservation plot for both layouts, respectively.

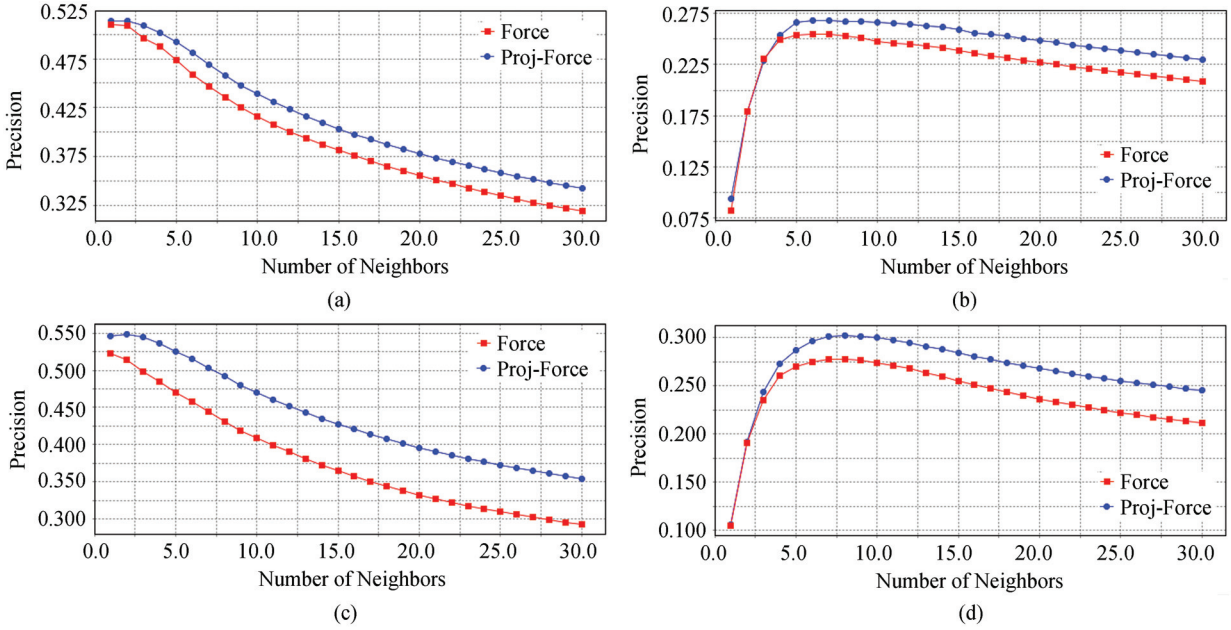

Fig.19. Precision plots for vis and agric datasets. (a) and (b) Neighborhood hit and neighborhood preservation for vis dataset, respectively. (c) and (d) Neighborhood hit and neighborhood preservation for agric dataset, respectively.

reason, besides the impossibility of reproducing neighborhoods for large graphs, is the boundaries between neighborhoods after averaging these values.

The plots in Figs. 18 and 19 show a prevalence of force-based placements after initial placement by IDMAP. Similar results were, however, obtained employing the projection technique LSP as initial placement.

For each of the line plots in those pictures, we also calculated the significance of the difference between *nhit*

and *npres* curves. We employed a simple statistical *t*-test, with resulting *p*-values for each plot above, shown in Table 4. They demonstrate the growing significance of the advantages of projection as an initial step as the number of nodes grows. The same table also shows the number of line crossings of the display, and the final number of groups used to calculate *nhit*. All these measurements confirm the advantages of node placement by multidimensional projection distribution prior to application of force-based placement.

4.3 User Evaluation

A user test was conducted with 30 users in order to verify the degree of difficulty relating attribute with relationships in our tool. Users were first-year undergraduate students with no knowledge on information visualization. None of the users knew the tool in advance.

Using the Orkut dataset, the test consisted in finding some relationship between the communities and their member attributes, first by coordinating the projection of the members and the projection of the communities, as seen in Figs. 15 and 16, and then using the heterogeneous network of members and communities, as seen in Fig.12.

A brief explanation (of about 10 minutes) of the test and basic functionality of the system was given at the start of the test. Then, the test was executed in three stages. At the start of each stage, the user initiated the filling of an answer form, at which point the starting time was recorded. At the end of each stage, the ending time was also recorded.

In the first test, users attempted to associate communities related to famous trademarks with gender and marital status using the coordinated views. In this test, all users could associate at least one trademark with gender and 11 users could associate at least one trademark with marital status. A total of 8 trademarks were freely found by users to be associated with gender or marital status.

In the second test, given a set of community names, the users attempted to discover whether these communities consisted predominantly of men, women, or none, using the heterogeneous network view.

Fig.20 shows a graph containing the results of the second stage of the test. As it is possible to see, the users could unanimously identify that communities 1 and 5 consisted predominantly of men (which they were), and 90% of the users could identify that community 2 consisted predominantly of women (which they also were).

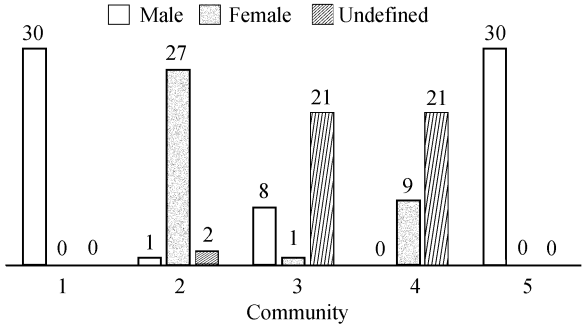

Fig.20. Graphic result of the second test. The bars indicate the number of users that classified the communities as being predominantly male, female, or undefined. Community 1 refers to "I love my Honda Civic", community 2 to "I seem snobbish, but I'm cool", community 3 to "I believe in love", community 4 to "No one can take away what is destined to be ours", and community 5 to "Car tuning".

For the final test, users were prompted to freely use the tool to find any other relationship between communities and members based on their attributes or associations, using the heterogeneous network view. They provided written answers of the associations found, and whether that association was what they expected.

Users found various new associations. Most of them could see that communities related to cars have many members in common, as well as communities related to feelings. However, the former is predominantly male, and the latter is predominantly female.

Additionally, they also found that communities related to drinks and games are predominantly male. Various users reported being surprised because the community "I believe in love" comprised both men and women, while they expected it would be predominantly female.

Subjects performed the test in a computing laboratory, where they stayed at most 45 minutes. The time to finish stages 1, 2 and 3 of the test were, in average, 11 minutes, 8 minutes and 7 minutes, respectively, suggesting that the tool allows fast identification of associations and quick learning.

Table 4. Comparison Values for Proj-Force and Force Plots

Dataset	*nhit p*-value	*npres p*-value	Edge Crossings Proj-Force	Edge Crossings Force	Number of Clusters
Eurovis	0.25	0.16	45 600	54 500	10
Vis	0.17	0.17	1 620 000	2 900 000	10
Agric	0.0013	0.029	4 000 000	7 900 000	15

Finally, at the end of the test, users were required to rate (0∼5) how easy, comfortable, and useful they found the tool. The results are shown in Fig.21.

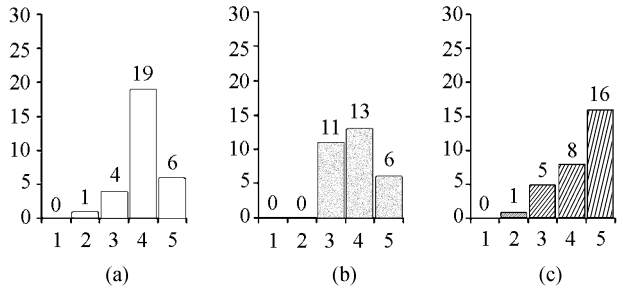

Fig.21. Ratings by users. (a) Ease of use. (b) Comfort. (c) Usefulness.

Users also classified the tool as useful to find interesting associations between members and communities, and they gave suggestions on how to improve the interface. A large media analysis company, that provided the Orkut data, also gave feedback on these solutions finding them useful.

4.4 Further Functionality and the PEx-Graph Tool

One way to employ PEx-Graph to analyze a social network is to build a visual display from a graph with nodes and edges, as well as their attributes, stored either in a VNA formatted file[33] or in a BibTex file[34]. The system has a wizard that supports defining specific layouts for the graph, based on connectivity, attribute, or simply a force-based layout from random initial points.

Users can generate views with or without edge information and coordinate variety of them together. There are also options to display content, that is, node attributes, remove nodes, split the graph and select regions.

Additionally, the system can calculate centrality measurements on the graph (such as degree, closeness, betweenness and clustering coefficient) and display them using visual attributes such as color or size of the nodes.

Nodes pointed by the mouse have their labels shown, and any of the attributes can serve as labels. Any display can be re-arranged by a simplified force-based algorithm subject to the weights of the edges (or values of distances when weights are not available), offering further views for exploration.

From the various functionalities available, two are worth highlighting. A very convenient functionality builds a co-authorship social network as well as a content-based projection of papers from a BibTex file. Co-author, co-paper and heterogeneous (author-paper) networks are built and, for the author-paper network, various sets of weighted edges are generated: author-author (weight is the number of papers in common), paper-paper (weight is the number of authors in common) and author-paper. A slider can decrease the number of edges depending on weight. Tools are provided to meet each of the requirements indicated in [3].

We have mentioned that multidimensional projections can be used and are suitable for analyzing various types of data with multiple attributes, from text to image collections. In PEx-graph, as illustrated before, it is possible to coordinate a view of the network with a view of individual data points projected using their content.

In the particular case where some or all of the nodes are text documents, users can explore, with the coordination function, document contents together with document networks, such as in a co-authorship social network.

Fig.22 gives an example of a heterogeneous co-authorship network having, in one side, a network of a small subset of papers and authors in the visualization field and, on the other side, a map of the article contents.

Similar papers are supposed to be mapped to neighboring regions using a previously available technique in the underlying software platform[25]. That type of multiple viewing of the dataset can be pre-assembled by the user in VNA and correspondence files, or automatically generated by PEx-Graph given any BibTex file.

Another tool that is bound to increase analysis performance is the sphere mapping available. In Fig.23, we show an example created in PEx-graph of the same co-authorship network, however, in this case, authors and papers are mapped onto the surface of a sphere, the former as rectangular boxes and the latter as smaller spheres.

The positioning algorithm is based on multidimensional scaling projection method in [35], where a stress measure is minimized. Stress is computed as the difference between the distances of points on the surface of the sphere and their original dissimilarities. Author attribute (in this picture, how many papers he or she co-authors) is represented with both color (from dark blue to red) and size of the nodes. Edge attribute (in this case, how many papers two authors have in common) is represented by color and thickness. Sphere mapping, besides offering an alternative of node distribution, can help augment the available space, therefore, decreasing clutter and line crossing.

The tool that implements these strategies, as well as supporting multidimensional visualization techniques, is made publicly available[36].

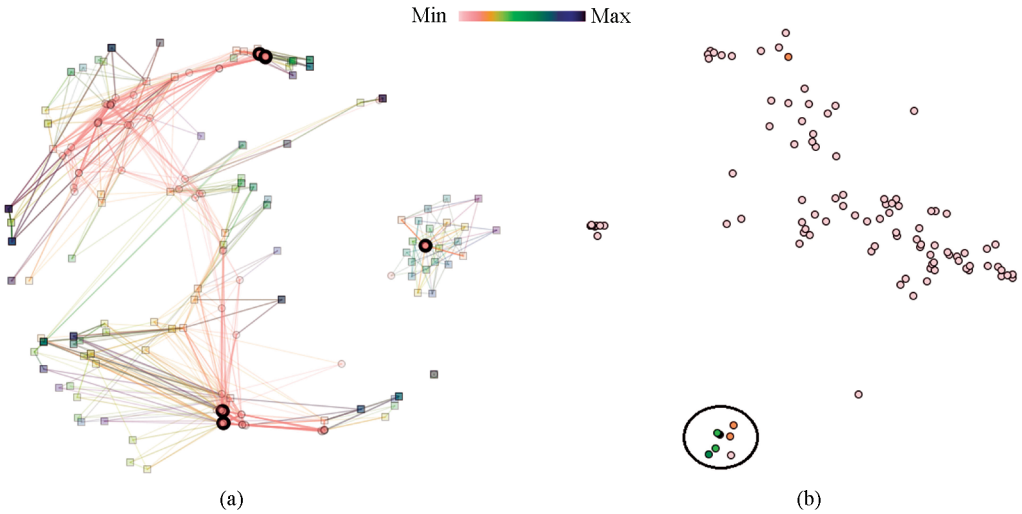

Fig.22. Network and content projections automatically built from a .bib file. Highlighted nodes in the network of papers (spheres) and authors (cubes) in (a) correspond to papers on graph visualization, circled in (b).

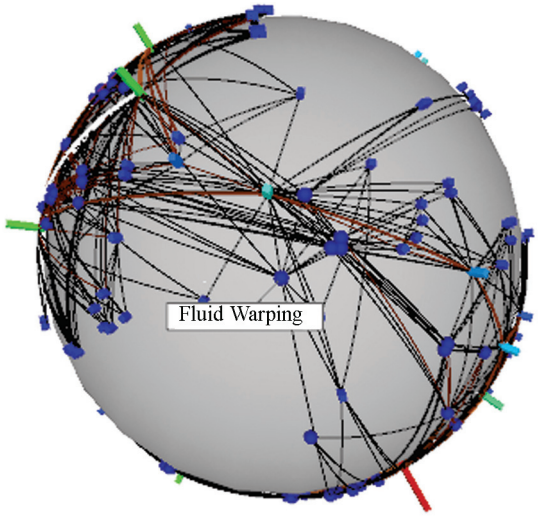

Fig.23. Sphere MDS Mapping of the co-authorship heterogeneous network.

5 Conclusions

The proposed approaches represent new alternatives for visual exploration of social networks. As suggested by some of the above examples, attribute layouts tend to place individuals with same attribute profile in the same region, while connectivity layouts tend to place individuals with similar connectivity distribution in the same region.

The results demonstrate that heterogeneous networks mapped using connectivity-based projections support query-driven exploration processes as well as general overview of the worlds within the network. When used to map nodes onto the visual plane prior to the execution of force-based layout, the achieved results are shown to present better node distribution when compared to conventional force-based graph drawing. There are fewer edge crossings, fewer force iterations are necessary to separate connected components on the layout, and node distribution according to their neighborhoods is also benefitted. We have compared the layouts numerically with support of simple statistical significance testing. The larger the graph, the better the benefit of pre-positioning by multidimensional projection.

Multidimensional projections also provide overview and detail views of communities and individuals based on their attributes and associations. Coordination helps the user associate both types of representation to find relationship patterns not directly expressed by either.

When considering a social network, key actors, as well as their roles and positions, must be identified. Embedded in how individuals connect to others and what the themes are of those connections is a large variety of indirect useful information. The techniques presented here support for revealing such embedded features. Regarding the roles and positions of the actors, the multidimensional projection approach groups individuals according to the similarity between their attributes.

By exploring the multidimensional projection, it is possible to see that individuals at the center belong to more communities, being similar to a large number of other individuals sharing some of their communities. Therefore, the positioning of the multidimensional projection reflects, to some extent, how an individual is central and how his or her themes relate, providing valuable information for various applications, such

as marketing, criminal investigation and knowledge domain analysis.

Although the projection utilized in most examples here is $O(n^2)$, it is faster than other algorithms concerned with finding subgraphs in otherwise cluttered views. Additionally, the approach decreases the problem of locality, which sometimes hampers force-based layout, impairing location of groups of interest. Two strong points of our approach are its support for focusing on regions of highly related entities and its immediate adaptation with other forms of content-based visual data analysis.

The software behind this initiative is made freely available[36] for general users.

Acknowledgments The authors are thankful to FAPESP, CNPq and CAPES for their financial support. They are also grateful to IBOPE Media and Innovation team for the Orkut data.

References

[1] Heer J, Boyd D. Vizster: Visualizing online social networks. In *Proc. IEEE Symposium on Information Visualization*, Minneapolis, MN, USA, Oct. 2005, pp.32-39.

[2] Huisman M, van Duijn M A J. Software for social network analysis. In *Models and Methods in Social Network Analysis*, Carrington P J, Scott J, Wasserman S (eds.), Cambridge University Press, 2005, pp.270-316.

[3] Henry N, Fekete J D. MatrixExplorer: A dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(5): 677-684.

[4] Henry N, Fekete J D, McGuffin M. NodeTrix: A hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 2007, 13(6): 1302-1309.

[5] Tulip Software. http://tulip.labri.fr/, 2011.

[6] Namata G M, Staats B, Getoor L, Shneiderman B. A dual-view approach to interactive network visualization. In *Proc. the 16th ACM Conference on Information and Knowledge Management*, Lisbon, Portugal, Nov. 2007, pp.939-942.

[7] Shen Z, Ma K L, Eliassi-Rad T. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(6): 1427-1439.

[8] Perer A, Shneiderman B. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(5): 693-700.

[9] Shneiderman B, Aris A. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(5): 733-740.

[10] Li C T, Lin S D. Egocentric information abstraction for heterogeneous social networks. In *Proc. International Conference on Advances in Social Network Analysis and Mining*, Athens, Greece, Jul. 2009, pp.255-260.

[11] Gloor P A, Krauss J, Nann S, Fischbach K, Schoder D. Web Science 2.0: Identifying trends through semantic social network analysis. In *Proc. International Conference on Computational Science and Engineering*, Vancouver, Canada, Aug. 2009, pp.215-222.

[12] Velardi P, Navigli R, Cucchiarelli A, D'Antonio F. A new content-based model for social network analysis. In *Proc. IEEE International Conference on Semantic Computing*, Santa Clara, CA, USA, Aug. 2008, pp.18-25.

[13] Bezerianos A, Chevalier F, Dragicevic P, Elmqvist N, Fekete J D. GraphDice: A system for exploring multivariate social networks. *Computer Graphics Forum*, 2010, 29(3): 863-872.

[14] Smith M, Giraud-Carrier C, Purser N. Implicit affinity networks and social capital. *Information Technology and Management*, 2009, 10(2-3): 123-134.

[15] Pretorius A J, van Wijk J J. Visual analysis of multivariate state transition graphs. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(5): 685-692.

[16] Archambault D, Munzner T, Auber D. GrouseFlocks: Steerable exploration of graph hierarchy space. *IEEE Transactions on Visualization and Computer Graphics*, Aug. 2008, 14(4): 900-913.

[17] Wattenberg M. Visual exploration of multivariate graphs. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, Montreal, Canada, April 2006, pp.811-819.

[18] Paulovich F V, Oliveira M C F, Minghim R. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proc. XX Brazilian Symposium on Computer Graphics and Image Processing*, Belo Horizonte, MG, Brazil, Oct. 2007, pp.27-36.

[19] Orkut. http://www.orkut.com/, 2011.

[20] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. *Communications of the ACM*, 1975, 18(11): 613-620.

[21] Minghim R, Paulovich F V, Lopes A A. Content-based text mapping using multi-dimensional projections for exploration of document collections. In *Proc. SPIE Visualization and Data Analysis*, San Jose, CA, USA, 2006.

[22] Navarro G. A guided tour to approximate string matching. *ACM Computing Surveys*, 2001, 33(1): 31-88.

[23] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: An International Journal*, 1988, 24(5): 513-523.

[24] Telles G P, Minghim R, Paulovich F V. Normalized compression distance for visual analysis of document collections. *Computers & Graphics*, 2007, 31(3): 327-337.

[25] Paulovich F V, Nonato L G, Minghim R, Levkowitz H. Least square projection: A fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 2008, 14(3): 564-575.

[26] Brandes U, Pich C. Eigensolver methods for progressive multidimensional scaling of large data. In *Lecture Notes in Computer Science 4372*, Kaufmann M, Wagner D (eds.), 2007, pp.42-53.

[27] Ingram S, Munzner T, Olano M. Glimmer: Multilevel MDS on the GPU. *IEEE Transactions on Visualization and Computer Graphics*, 2009, 15(2): 249-261.

[28] Jolliffe I. Principal Component Analysis. New York, NY, USA: Springer, 2002, p.487.

[29] Netlog. http://www.netlog.com/, 2011.

[30] Fruchterman T M J, Reingold E M. Graph drawing by force-directed placement. *Software —Practice & Experience*, Nov. 1991, 21(11): 1129-1164.

[31] Zhang Z K, Zhou T, Zhang Y C. Tag-aware recommender systems: A state-of-the-art survey. *Journal of Computer Science and Technology*, 2011, 26(5): 767-777.

[32] The Collection of Computer Science Bibliographies. http://liinwww.ira.uka.de/bibliography/, 2011.

[33] Analytic Technologies. http://www.analytictech.com/netdraw/netdraw.htm, 2011.

[34] BibTeX. http://www.bibtex.org/, 2011.

[35] Cox T F, Cox A A M. Multidimensional scaling on a sphere. *Communications in Statistics — Theory and Methods*, 1991, 20(9): 2943-2953.

[36] VisGraph. http://infoserver.lcad.icmc.usp.br/infovis2/Tools, 2011.

Rafael Messias Martins is currently a Ph.D. candidate at University of São Paulo, São Carlos, Brazil. He received his B.Sc. and M.Sc. degrees from University of São Paulo, Brazil. His main research interests include information visualization and software engineering.

Gabriel Faria Andery is currently a software developer at CPM Braxis Capgemini, São Paulo, Brazil. He received his B.Sc. degree in computer science in 2008 and his M.Sc. degree in 2010 both from University of São Paulo (USP), in São Carlos, Brazil. His research interests include information visualization and complex networks.

Henry Heberle received his B.Sc. degree in computer science from University of São Paulo, São Carlos, Brazil. His research interests include information visualization and network analysis.

Fernando Vieira Paulovich obtained his B.Sc and M.Sc degrees in computer science in 2000 and 2003 respectively from Federal University of São Carlos, São Carlos-SP, Brazil, and received his Ph.D. degree in computer science in 2008 from University of São Paulo, São Carlos-SP, Brazil. His main fields of interest are information visualization, visual data mining and visual analytics. Currently, he is a lecturer and researcher at University of São Paulo, Brazil.

Alneu de Andrade Lopes holds a Ph.D. degree in computer science from University of Porto, Portugal. Currently, he is an assistant professor at University of São Paulo, Brazil. His main research interests include artificial intelligence, machine learning and data mining.

Helio Pedrini received his Ph.D. degree in electrical and computer engineering from Rensselaer Polytechnic Institute, Troy, NY, USA. He received his M.Sc. degree in electrical engineering and his B.Sc. in computer science, both from the University of Campinas, Brazil. He is currently a professor in the Institute of Computing at the University of Campinas, Brazil. His research interests include image processing, computer vision, pattern recognition, computer graphics, computational geometry.

Rosane Minghim is an associate professor at University of São Paulo, São Carlos, Brazil. She is interested in all aspects of visualization, information visualization and visual analytics.