EG UK Theory and Practice of Computer Graphics (2009)
Wen Tang, John Collomosse (Editors)

Visual Representation of Multiple Associations in Data using Constrained Graph Layout

W. Xu and J. Sreevalsan-Nair

Texas Advanced Computing Center, The University of Texas at Austin, U.S.A.

Abstract

This paper presents a new approach for simultaneously visualizing multiple exclusive associations, defined on the same dataset, using constrained graph layout. We work with two different associations at a time, which can be represented as a binary graph operation. Given an initial graph layout which represents an association of the data, another set of constraints is applied to the graph to represent a second association on the same dataset to obtain the final layout. Our motivation is to preserve some features of the first layout as well as to achieve a simultaneous view of both the associations. We use this approach generically for three applications: for visualization of data with geometric and categorical constraints, respectively. We further propose to extend it to multiple associations, by using the binary operation multiple times.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications

1. Introduction

Our motivation is to develop a graph layout to visualize multiple associations in datasets simultaneously. In the graph visualization community, several applications inherently achieve this need in a subtle manner - for instance, graph layout of spatio-temporal datasets and cluster visualization depict multiple associations simultaneously. However quite often these applications cater to specific needs, and our goal is to use one generic approach for such various applications.

In this paper, we have implemented our approach for applications with two different associations for the same dataset, and we propose to extend it to more than two associations in any given dataset. The datasets on which we have applied our algorithm to, either have geometric constraints or categorical information for clustering. Our focus is on applications where the data can be depicted as a simple graph layout using an association in the data and the second association can be represented as a set of constraints to modify the first layout using spring force model. Our applications include (i) datasets with geometric constraints, which include prediction of interaction pairs in RNA sequence data and a time-series of a motion sensor network (ii) datasets with categorical information, such as clustering in a dataset of email communications in Enron. For both these sets of applications, the position of the vertices are strictly enforced by the constraints. In our applications with geometric constraints, the various sensors are placed in predetermined locations in the building, and the nucleotides are placed in specific positions in the secondary structure of a RNA sequence. In applications with categorical constraints, there is an additional step to convert the constraints to geometric ones.

2. Related Work

Graph-drawing techniques and information visualization using graphs are widely-researched topics. Dynamic graphs are used for representing time-varying data. In [BW97], dynamic graphs are created by initially creating a mental map and then retaining the mental map partially in consecutive layouts. In [FT08], the dynamic graph layout is achieved by using force-directed methods [KW01] and representing it on GPU for real-time generation of large graphs. More work on dynamic graph layout using force-directed methods can be found in [DG02, BFP05].

Several variants of standard algorithms for clustering using graph layout have been designed for specific results. Using dummy vertices and edges in a spring force model with differentiation of intra-cluster and inter-cluster forces gives a successful cluster layout [HE98]. Though in a sim-

DOI: 10.2312/LocalChapterEvents/TPCG/TPCG09/065-068

ilar vein, our approach is not limited to convex bounding boxes, and does not need a cluster-tree structure to aid in the layout. Qian et. al. [QZL04] used constraint-based clustering combined with graph partitioning, and min-max principle. Dwyer et. al. [DM06] used hierarchical graph layout and specific band constraints to determine clusters.

In [AAPS05], the vertices in the graph are in specific geographical regions, and their mobility is limited by certain constraints imposed by underlying geography. Though there is similarity in our and their work with respect to confining vertices within bounding boxes, the scope of their method is to find the optimal constrained graph layout with aesthetic criteria, while ours is directed towards a generic method for simultaneous visualization of multiple associations.

Similar to our applications, there is a lot of work in simultaneous visualizations using composite views. Ivanov et.al [IWSK07] developed a visualization using multiple views to simultaneously show animation through time, using time-line controls to show local and global temporal changes. Simultaneous graph layout [EKLN05] shows three different visualization schemes to view two different relationships of the same set of vertices in a graph. Different from our approach, they independently preserve the two graph layouts representing the relationships.

3. Our Approach

Our graph-drawing algorithm is essentially a binary graph operation - when applied to a dataset with two associations, we represent the initial association as a graph and the second one is then applied as a set of constraints to the initial layout to obtain a final layout. The second association can also be represented as a graph if it contains specific geometric information. The requirement of the second association to be a geometric constraint can be met easily - if the second association is a geometric constraint, we use the constraints explicitly; and if it is a non-geometric constraint, such as categorical constraints, we implicitly reduce them to geometric constraints. Thus our approach takes in two graph layouts to give an output layout, just as in a binary graph operation.

Our algorithm, explained as pseudo-code in Algorithm 1, takes in a dataset D(R_1, R_2), where R_1 and R_2 are two different associations in D. We reduce this to an undirected planar graph $G(V,E,C)$, where V, E, and C are the sets of vertices, of edges connecting vertices in V, and of constraints on V or E, respectively. Each constraint in C on vertices in V may be in the form of a geometric bounding area, B_a or containment in a cluster or category c, which is further converted to a virtual bounding area B_a which is physically represented by the set of the vertices belonging to c. We use dummy vertices and edges to apply the constraints, similar to [HE98]. For the final graph layout, we use attractive intra-cluster and repelling inter-cluster forces on a force-spring model [KW01], which gives the following effects: (i) the vertices lie within their respective bounding areas (ii) the bounding areas move away from each other and converge to a non-overlapping state.

Algorithm 1 Create a constrained graph layout simultaneously displaying two associations.

Require: D(R_1,R_2)
  Derive V and E from R_1 and C from R_2
  Layout V and E using a graph layout G
  if constraints in C are categorical **then**
    for each category c in C **do**
      Use the set of $\forall v \in$ V belonging to c as a virtual bounding area, B_a
    end for
  else
    Use bounding area B_a represented by each geometrical constraint in C
  end if
  for each bounding area B_a defined in C **do**
    Find centroid v_c of all $v \in B_a$
    Connect v_c to $\forall v \in B_a$ using dummy edges $e \in E_d$, representing intra-cluster forces.
  end for
  Link all representative vertices, $\{v_c\}$, using dummy edges $e \in E_d$ to give a fully connected graph, representing inter-cluster forces.
  Re-layout $G(V+\{v_c\}, E+E_d)$ using force-spring model.

Our work is motivated as an extension of the interactive visualization for relational databases [XG08]. The graph layouts of our test datasets, which are retrieved from relational databases, are rendered using a Java implementation using prefuse [HCL05].

4. Preliminary Results: Applications

4.1. Visualization of Data with Geometric Constraints

We explore the application of our approach to datasets with geometric constraints, which are the secondary associations applied on the data. The data is represented by an initial graph layout defined by an inherent association. In the following examples, by an infinitesimally sized bounding area, we refer to a bounding box of size 1.1% of vertex size, centered at the vertex.

Visual Inspection of Base Pair Interaction Prediction in RNA Sequence Data

Biological sequences have natural forms as three dimensional structures which can be mapped to a two dimensional space known as the secondary structures, which are often illustrated using secondary structure diagrams. Although nucleotide sequences are often represented as one dimensional strings, it is beneficial to visualize these sequences in conjunction with their structural information. Figure 1 shows

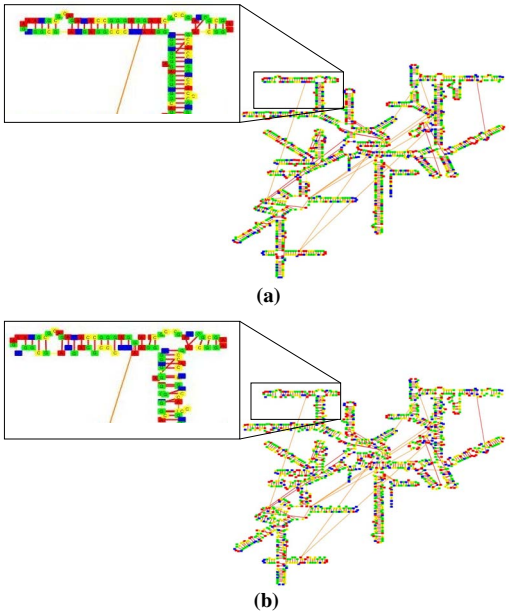

Figure 1: *Graph layout of RNA sequence data generated using our approach, with (a) with strict constraints and (b) infinitesimal bounding box constraints, on the vertices. The nucleotides in RNA sequence data are represented as vertices and the edges represent the known and predicted interactions between base pairs in the sequence. The vertex color indicates nucleotide type.*

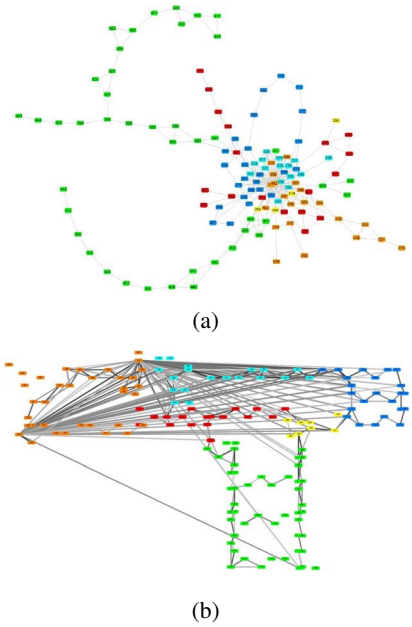

Figure 2: *Graph layout of a month-long activations of motion sensors in the MERL building using (a) force-directed graph layout, and (b) our approach with infinitesimal bounding box constraints. The sensors are represented as vertices and the edges represent 1000 or more concurrent activations between two given sensors over time. The vertex color represents the different wings of the building and the edge color represent the frequency of the concurrent activations.*

secondary structure of 16S rRNA sequence of *Thermus thermophilus* as a graph layout whose vertices and edges represent nucleotides and predicted interactions between pairs of nucleotides, respectively.

The predictions of these interactions are based on covariation analysis of the RNA alignment. The motivation for visualizing these base pairings of RNA sequence data is to get an insight on the quality of the predictions. This visual analysis helps us in assessing a probability score giving the likelihood of interaction between two given nucleotides.

Time-series of Activations in a Motion Sensor Network

The motion-sensor data was recorded from a network of 215 motion sensors over the course of about a year in two floors of MERL office-space [IWSK07, WILW07]. The sensors report the presence of motion in their field of view and stores the time stamp of the activation in the database.

We generate a planar graph using the data from one of the two floors of the building. The sensors are represented as the vertices of the graph and the edges show concurrent activations between any two sensors. The color of the edges represent the frequency of such concurrent activations - the higher the frequency the lower the gray-scale of the color. We compare the graph layout using force-directed method and our approach in Figures 2(a) and (b), respectively. In the latter we restrict the sensor positions to an infinitesimal bounding boxes around their physical locations. We can see that in our approach, in addition to preserving the physical layout of the sensors, the pairs of sensors with more frequent consecutive activations move towards each other.

4.2. Visualization of Data with Categorical Constraints

Our approach is applicable to visualization of two different relationships among datasets without explicitly defined geometric constraints, as in the case of categorical constraints. Multiple associations can be derived within the same dataset and the secondary associations can be represented using implicit geometric constraints, as explained in Section 3.

Categorization in Enron email Database

The Enron email database comprises of a social network involving 87,474 employees and 255,636 email communications between them, which are uniquely classified into 13 categories. The data used here is a subset of the emails sent by persons with id 36 and 242, respectively. This subset of emails falls into nine different categories. For each category,

a constraint can be implicitly defined and the nodes from the same category are regrouped together using constrained layout, as shown in Figure 3. Our approach gives a successful clustering of the different categories of the emails.

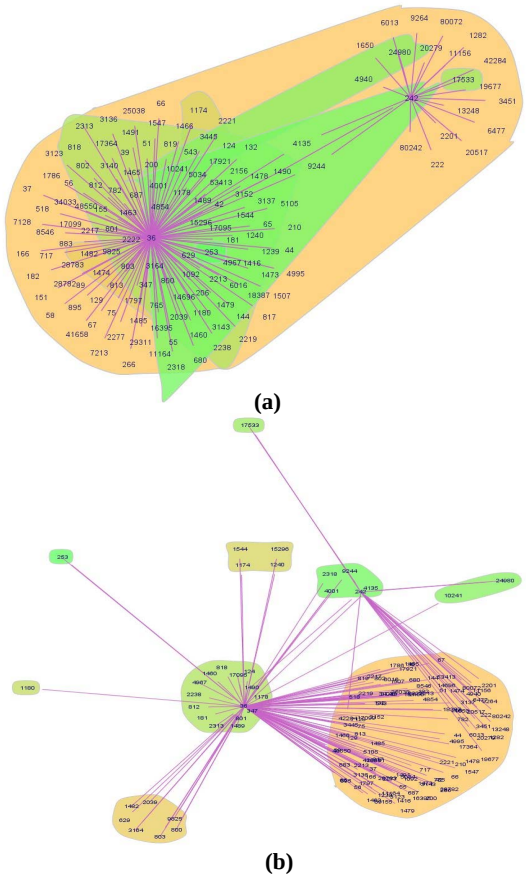

Figure 3: *Comparison of graph layouts with categorical constraints, in a subset of Enron email dataset Network graph (a) before and (b) after applying categorical constraints, where categories of nodes are differentiated by color shading.*

5. Discussions & Future Work

Scientific datasets commonly consist of multi-dimensional information. Thus deriving multiple associations from the same dataset by combining various aspects of data is plausible and such an exercise enables us to identify various patterns in the dataset. For example, the RNA sequence visualization in Section 4.1 is motivated by the need of a biologist to visually evaluate results of interaction predictions. This visual representation has helped in further research questions, such as possibility of clusters of predictions. Our approach successfully explores a composite representation of two associations of a dataset using constrained graph layout.

Several standard visualization techniques, such as color mapping, glyph representation, etc., can be inherently represented as additional associations, which can be composited into the graph layout thus extending our approach to more than two associations. We propose to use the binary graph operation multiple times compositing two associations at a time to achieve composition of multiple associations. A good research question to pursue is the associativity of compositing multiple associations. Enabling flexibility of inclusiveness of multiple associations either conditionally or user-defined is another direction to pursue. Our future work will focus on improving the layout algorithm and enabling dynamic constrained graph layout. Currently, various parameters for the layout algorithm, such as spring force coefficient, spring length, etc., are predetermined proportional to the category size, for cluster visualization. This leads to small clusters being placed far apart, as can be seen in Figure 3. We expect that a heuristic-based algorithm will be able to resolve issues with determining parameters for the layout algorithm.

6. Acknowledgements

The authors are grateful to a TeraGrid Resource Provider Grant for funding and the Visualization & Data Analysis Group, TACC for supporting this work. The RNA secondary structure file, the MERL sensor data and the Enron email database are from http://www.rna.ccbb.utexas.edu/, IEEE Infovis Contest 2008 (http://www.merl.com/wmd/infovis.html), and http://bailando.sims.berkeley.edu/enron_email.html, respectively.

References

[AAPS05] ABELLANAS M., AIELLO A., PENALVER G. H., SILVEIRA R. I.: Network drawing with geographical constraints on vertices. *Actas XI Encuentros de Geometriá Computacional* (2005), 111–118.

[BFP05] BRANDES U., FLEISCHER D., PUPPE T.: Dynamic spectral layout of small worlds. In *Graph Drawing* (2005), pp. 25–36.

[BW97] BRANDES U., WAGNER D.: A bayesian paradigm for dynamic graph layout. In *GD '97: Proceedings of the 5th International Symposium on Graph Drawing* (London, UK, 1997), Springer-Verlag, pp. 236–247.

[DG02] DIEHL S., GÖRG C.: Graphs, they are changing. In *GD '02: Revised Papers from the 10th International Symposium on Graph Drawing* (London, UK, 2002), Springer-Verlag, pp. 23–30.

[DM06] DWYER T., MARRIOTT K.: IPSep-CoLa: An incremental procedure for separation constraint layout of graphs. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 821–828.

[EKLN05] ERTEN C., KOBOUROV S. G., LE V., NAVABI A.: Simultaneous graph drawing: Layout algorithms and visualization schemes. *Journal of Graph Algorithms and Applications 9*, 1 (2005), 165–182.

[FT08] FRISHMAN Y., TAL A.: Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics 14*, 4 (2008), 727–740.

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: A toolkit for interactive information visualization. In *Conference on Human Factors in Computing Systems* (Portland, OR, USA, 2005), ACM Press, New York, pp. 421–430.

[HE98] HUANG M. L., EADES P.: A fully animated interactive system for clustering and navigating huge graphs. In *GD '98: Proceedings of the 6th International Symposium on Graph Drawing* (London, UK, 1998), Springer-Verlag, pp. 374–383.

[IWSK07] IVANOV Y., WREN C., SOROKIN A., KAUR I.: Visualizing the history of living spaces. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov./Dec. 2007), 1153–1160.

[KW01] KAUFMANN M., WAGNER D. (Eds.): *Drawing graphs: methods and models*. Springer-Verlag, London, UK, 2001.

[QZL04] QIAN Y., ZHANG K., LAI W.: Constraint-based graph clustering through node sequencing and partitioning. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD* (2004), pp. 41–51.

[WILW07] WREN C. R., IVANOV Y. A., LEIGH D., WESTHUES J.: The merl motion detector dataset. In *MD '07: Proceedings of the 2007 workshop on Massive datasets* (New York, NY, USA, 2007), ACM, pp. 10–14.

[XG08] XU W., GAITHER K. P.: On interactive visualization with relational database. *Compendium of IEEE Visuallzaiton 2008* (2008), 116–117.