LinkScope: A Toolkit for Interactive Graph Analysis

Yun Teng*
Dept. of Computer Science,
University of California, Santa Barbara.

Tobias Höllerer†
Dept. of Computer Science,
University of California, Santa Barbara.

John O'Donovan‡
Dept. of Computer Science,
University of California, Santa Barbara.

ABSTRACT

This paper presents LinkScope, a toolkit for interactive analysis of node link graphs that supports dynamic addition of attributes from tabular data. The interaction technique draws on ideas from 3D modeling, mesh deformation, and static graph drawing to promote discovery of hidden information across a wide variety of graph types and analysis tasks. The key innovation of this work is the application of methods traditionally reserved for automated graph layout and clustering, to produce useful task-specific layout through dynamic interactions. Graph nodes are dynamically repositioned using an interpolated decay function over a single node movement provided by a user. We describe several variants of the interpolation method, including coupling it with a fast local-cut algorithm for cluster selection. Compared to traditional layout mechanisms the technique is particularly useful when meta-data nodes are added to a graph, increasing its connectivity. We show how the techniques can be used interactively to solve analysis tasks based on graph connectivity alone, and illustrate further benefits when coupled with content-based analysis and text-based search. Validation of both approaches is shown through use cases that highlight unique analytical benefits of each technique on a variety of tasks. Our use cases cover a collection of 16K awarded NSF grant proposals with meta-data, a corpus of New York Times news articles and a collection of 7000 incident reports about the historical conflict in Northern Ireland.

Index Terms: H2.8 [Database Management]: Interactive Data Exploration/Discovery—Data and Knowledge Visualization
H5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical User Interfaces

1 INTRODUCTION

Improvements in both hardware and software technologies are making interactive visual analytics increasingly more suitable as a solution to information overload. The key contribution of the social web –user generated content, is being produced from an abundance of sources, such as social networking applications, blogs, wikis, microblogs and digital libraries, to name a few. Many solutions to information overload have been tried and tested over the years, from the familiar index-based search engines to content-based and collaborative filters in recommender systems, collaborative search, social bookmarking services, to more structured, semantic solutions such as ontology-based search in SPARQL endpoints. All of these tools and techniques filter content in some way or other, whether passively though preference modeling, or actively through user-specified search queries. One factor that prevails is that the resulting refined information space can vary greatly in terms of content items and the interconnections that may or may not exist between them. Consider a tabular database of awarded NSF research grants for example: without advance knowledge of the database, an analyst who issues a query for "NSF program officers who awarded large grants to Californian institutions in the last year" can receive a network (of people and institutions) of arbitrary size and connectedness.

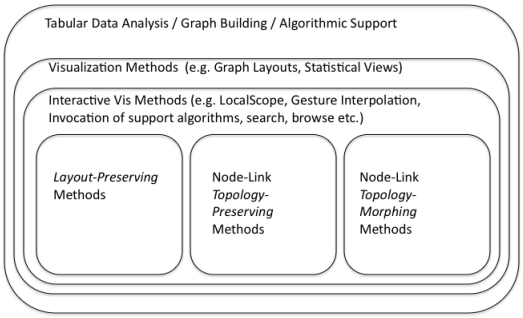


Figure 1: Abstract stratification of the analysis tools available in LinkScope.

In this paper, we aim to explore this issue of uncertainty about scale and connectivity properties in filtered result sets, in particular as they apply to the process of visual analytics. To navigate a complex information space for a particular task, an analyst may construct many different types of node-link networks from an underlying data source, or set of sources, for example, by selecting attributes in a tabular structure to link entities together for a visual representation, or by adding and removing nodes and edges based on some relevance metric. This leads to an uncertainty in the complexity of the graph to be visualized for further analysis. Traditional graph layout mechanisms have inherent limitations which can render them unsuitable for some of these networks, typically as a result of scale or connectivity limitations [9].

To address this issue, we introduce a lightweight visual analytics framework which aims to support the key steps in visual data exploration: Data gathering, modeling, visualization, exploration and insight. [22]. The goal of the framework is to test the utility of two novel techniques for the interactive manipulation of node-link graphs. Both techniques are designed to work in tandem with existing graph layout algorithms, to help analysts make sense of result graphs, particularly when they become too complicated for a standalone layout algorithm. Figure 1 presents a high level classification of the data analysis processes in the framework. Analysts can manipulate raw data in tabular form to construct graphs by selecting various column headers to link data entities. This enables the analyst to construct a broad variety of node-link graphs tailored to a given analysis task. Filters can be also applied based on attribute types in the data to further refine the graph. Once a graph has been built, automated force-based algorithms can be applied to produce graph layouts. A set of statistical analyses tools and visualizations can be used in parallel to help the analyst understand more

*e-mail: yunteng.cs.@cs.ucsb.edu
†e-mail:martha.stewart@marthastewart.com
‡e-mail:ed.grimley@aol.com

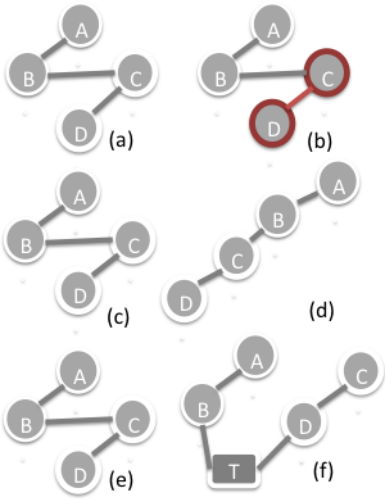


Figure 2: Example interactions in each of three classes a) to b) shows layout-preserved, c) to d) shows topology preserved, and e) to f) shows an interaction that does not preserve topology, such as the splicing of a meta-data node between two of the original nodes.

complicated graphs, for example by examining link distributions and clusters. The primary focus of this work is on the next class in Figure 1: Interactive visualization methods. For the purpose of discussion and testing, we further classify this into three core areas for node-link graph manipulation, as illustrated by the node-link diagrams in Figure 2 1) Methods that preserve layout, 2) methods that preserve topology (node-link structure), and 3) methods that change the underlying node-link structure in some way.

2 Contributions

The contributions of this paper to interactive visual analytics can be organized as follows:

- A lightweight, open-source and web-based framework for performing visual analysis of tabular, node-link and/or text-based data.
- Two novel techniques for interaction with node-link graphs. One based on interpolation of mouse gestures over the graph, and a second technique that applies EvoCut [8] conductance-based clustering and other methods prior to gesture interpolation.
- An evaluation of the interactive analysis capabilities of the framework and the interaction techniques on a set of NYT news article, a corpus of awarded NSF grants from 2008 to 2010, and on a database of terrorism-related deaths during the Northern Ireland "troubles" from 1972 to 1985.

The remainder of this paper is organized as follows. Firstly, we will present an overview of the analysis toolkit, using an example work flow. This includes a description of the available tools and algorithms in the toolkit, organized using our earlier classification from Figure 1. The following section ends with a discussion of the scope of different workflows that are available. Next, we describe our two novel graph interaction algorithms in detail, including a discussion of computational complexity for each. Finally we evaluate the analysis toolkit with particular focus on the role of the interaction algorithms in the three usage scenarios.

3 LinkScope Toolkit

Figure 3 shows a snapshot of the interface during an example analysis session. In this session, an NSF executive is interested in finding out the various ways the foundation is funding work related to counter-terrorism. The data shown is a set of awarded NSF grant documents from 2008 to 2010 inclusive. The main window shows the current graph view. On the left of the screen is a provenance view, where the smaller graphs depict the state of the visualization at each preceding filtering step. Each view can be reverted to at the click of a button. The fan-like panels on the right control a variety of data mining and visualization algorithms which can take selected data from the graph view as input.

The top left view shows the initial graph of 6000+ connected entities from the NSF data set. Most of these nodes represent grant documents, and they are connected through a small percentage of "topic" nodes. These nodes have been computed using Latent Dirichlet Analysis over the contents of each document, using the algorithm outlined in [11]. The layout has been produced by running a simple multidimensional scaling over the set of inter-topic similarity values output by the algorithm, thereby placing similar topic nodes in close proximity to each other. Each document/grant node has been colored based on its NSF category. For example, network science, graphics, cyber-trust, visualization etc. The clusters of color arise from the inherent topic similarity across the set of grant documents. We believe that this provides a good overview reference to begin our analysis workflow.

Next, a simple text-based query was issued over all documents for the keyword "counter-terrorism". For simplicity in this example, the search was limited to document abstracts only. The search returned a list of ten documents in a panel on the right. The analyst checks all of them and they appear on the graph as isolated nodes. At this point, a number of expand algorithms can be run to search the neighborhood of these nodes in the topic graph. In this case, the analyst is interested in a more targeted search, so, using the "tabular data" panel on the right, she opts to build edges based on the AWARDEE_STATE property. This results in the second view, which is a set of largely disconnected clusters, but with the addition of new nodes on the graph to represent the connecting attribute. The analyst can optionally apply filters on the connecting attributes at this point. For example, to view only those grants related to CA and VA states. A list of active filters is shown at the bottom of the panel. After each addition to the graph, a layout algorithm (in this example a simple Fruchterman-reingold algorithm, but optionally an FM3 [20], or binary stress layout) is automatically invoked to produce a clearer view of the new graph. The analyst repeats this process by adding edges for PROGRAM_OFFICER and COUNTRY, and the graph becomes much more connected, as shown in the remaining provenance views on the left. Edges and filters can be applied for any attribute or attribute value in the data set.

In the last provenance view, the graph has already become fairly connected, and the force directed layout is already losing clarity due to crossed edges etc. At this point, the analyst notices a seemingly prominent node (Leyland M. Jameson), and would like to investigate further. The analyst selects the node and drags it in an arbitrary direction, automatically invoking an interpolation algorithm which applies the gesture to all other nodes in the graph with a decaying weight based on hop distance from the moved node. For visual landmarking, a smooth animation is applied over the node transitions, and the graph morphs into the view shown in the main window. This tree-like structure pivots nodes around the target, and provides a clearer view of its relations. For instance, Leyland M Jameson was program officer on NSF grants related to counter-terrorism in three states: California, Texas and Michigan.

Now that we have provided a high level overview of the LinkScope toolkit, we discuss is feature set more thoroughly. The next paragraphs organize the available tools based on our interac-

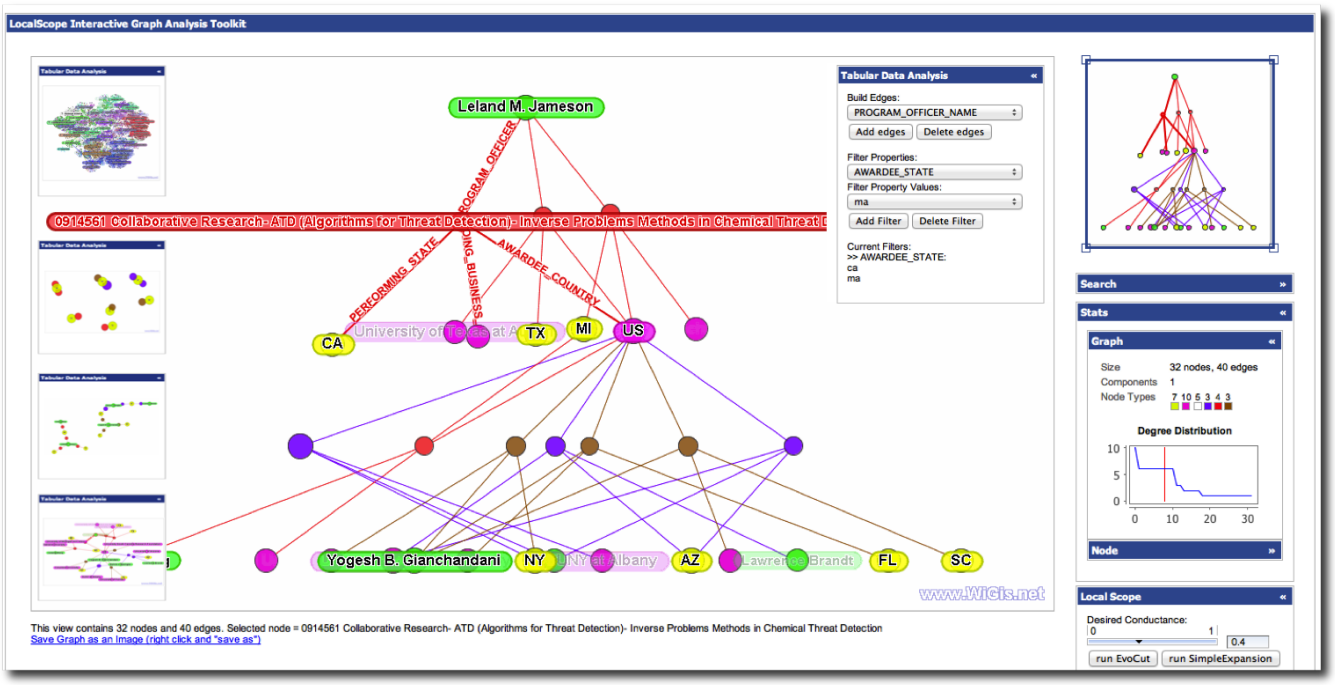


Figure 3: Example workflow in LinkScope, showing an interpolated layout of filtered NSF data.

tion types from Figure 2 earlier. Following this, a detailed description is provided for each graph interaction technique. LinkScope is implemented using a Java-based graph visualization framework called WiGis [17], which provides it with a standard set of interactions such as zoom, labeling, appearance adjustment etc. Further details of these are available in [17].

3.1 Graph Construction and Layout

Prior to visualization, the analyst can build a graph from one or more data tables by building edges from attribute types and/or filtering edge sets based on attribute values. Liu [24] describe a set of detailed mechanisms in for building graphs from tabular data, such as weighting, projection and aggregation, –all of which would be useful additions to our framework. In this paper we are interested primarily in mouse-based gesture mechanisms for graph interaction, and accordingly the discussion of tabular data ingestion here is relatively brief.

Edge Builder LinkScope currently ingests tabular data by treating each row as an entity and each column header as an attribute for that entity. The initial graph contains a set of disconnected entities, each representing a row in the tabular data. The tab panel shown in Figure 3 allows an analyst to build a custom graph by adding attribute nodes relevant to her current task. Attribute nodes have labeled edges linking to entity nodes (NSF grants and Terrorism Incidents in the use cases presented later). Figure 3 shows an example of type labeling along attribute edges.

Edge Filter Obviously the addition of attributes as graph nodes quickly increases graph connectivity leading to the classic "ball of string" problem. An edge filtering feature partially alleviates this by supporting visual examination of arbitrary subsets of nodes with particular attribute values. While there are other useful ways to perform this selection, for example, binning, pivoting are proximity grouping [24], the building and filtering process is sufficient to support our analysis of the usefulness of our interaction techniques. Furthermore, LinkScope supports a window-based filtering over ordered attributes such as time, as described next.

Time-based Slicing Temporal distribution of events is critical for many analysis tasks. For example, [27] highlight the usefulness of temporal distribution by analyzing the hash tag #earthquake in twitter streams to pinpoint each aftershock in the 2010 Chilean earthquake. LinkScope supports visual slicing of ordered attributes such as time by running a layout on nodes within a specific time window. Analysts can click on a "play" button to step through the current window, incrementally adding nodes with the next time stamp to the visualization, based off a pre-computed layout for the entire time window.

Topic Modeling LinkScope uses libraries from the TopicNets system [17] to perform LDA topic modeling on a graph prior to visualization in order to link nodes based on the topic similarity of their content. This feature allows the system to infer inter-node similarity based on the topic associations, enabling it to work on free text documents with no structured meta data.

3.2 Automated Visualization Methods

Once an initial graph has been built, LinkScope can apply a range of "non-interactive" algorithms to produce various perspectives on the graph. Here, we define "non-interactive" as the set of algorithms where the analyst does not click on the graph itself. Most of the methods below need to be invoked via controls on the right panel, while the results are output directly on the node-link graph.

Content Highlighting A simple, but very useful tool is the text-based search panel. Analysts can search over various abstractions of node contents, such as "Label", "Abstract" or "Full Contents" in our NSF scenario for instance. Search results appear in a list and can be further filtered via check-boxes in the search panel. Results are then highlighted on the graph in the main window.

Clustering To provide an abstract view of the node-link topology of a graph, simple k-Means clustering is supported through a control on the right panel. The graph maintains a layered hierarchy so that an analyst can view graph connectivity at different levels of granularity using a slider. An additional feature to the clustering

algorithm is that mouse interactions with a clustered, abstract view of a graph are mapped onto the full graph, and an analyst can view these in a window on the right panel.

Statistical View Statistical views have proven to be a useful analytical aid for graph visualization [26, 21, 29, 18], especially for larger, unfiltered and more highly connected graphs. LinkScope supports a set of three panels that provide simple statistical data insights about a currently selected subset of the graph. The first panel, shown on the right of Figure 3 provides basic stats at the global level. A degree distribution graph plots the number of nodes with a given degree, and overlays the currently selected node(s) as vertical red lines, indicating their position in the global distribution. A color-coded list shows the number of and size of disconnected components in the graph. This panel is interactive, and clicking on a distribution or component value will highlight the appropriate node in the main view. The second panel (not shown) computes pairwise statistics (such as shortest path for example) and is only displayed when exactly two nodes are selected. This supports automatic drawing of the shortest path between node selections on any visual configuration in the main window. Lastly, the node-level statistics panel (shown closed in Figure 3 provides detail about individual nodes/entities, for example, an analyst click though to the original entity/document that it represents. The panel shows a list of neighboring nodes, and supports drill-down for those also.

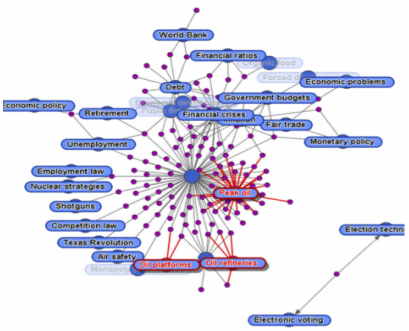
Layouts The LinkScope analysis toolkit contains implementations of a number of force-directed graph layout algorithms including an optimized Fruchterman Reingold layout [17] Binary stress layout [23], FM3 layout [20] and the topic-based multidimensional scaling layout from [14]. The toolkit also uses layered layout based on Dk component-analysis [25] that attempts to lay out more connected components first.

4 Interactive Graph Analysis

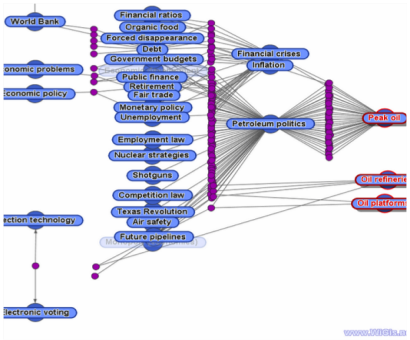
In this section, we describe our novel interaction algorithm in LinkScope, including a discussion of various design choices and implementation detail. The idea is based on our original interpolation method in [35]. The technique has been extended to support untangling highly connected graphs while preserving informative local structures. To clearly illustrate the effect produced with each technique, Figures 4 and 5 show the results on a simple bipartite graph. The nodes in the graph represent a set of news articles crawled from the New York Times website. A topic modeling algorithm [11] has been run over the contents of the articles, and additional nodes have been added to represent the resultant topics (i.e.: the term lists produced by the topic modeling algorithm). Edges are placed between documents and topic nodes if they have an association score above a threshold value.[14] provides further detail on generating document-topic graphs in this manner. Figure 4(a) shows an overview of this graph laid out with a simple force directed algorithm. Additionally, a text based search over the node contents has been performed for the keyword "oil" and the result nodes have been highlighted in red. The blue labeled nodes represent mined topics and the smaller red nodes represent individual news articles.

4.1 Interpolation method

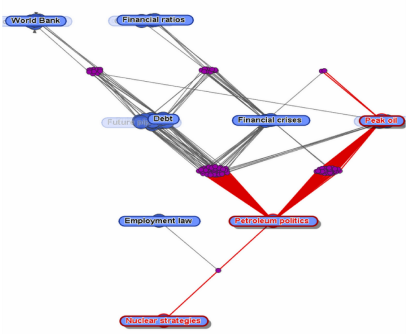
The original interpolation method is a simple approach to interactive graph layout. The algorithm is instantiated with one single node that the user clicks on. The displacement made on this target node is applied to all the other nodes with a weighted function that decays based on hop-distance from the moved node. This will deform the graph in a tree-like fashion. This method allows an analyst to "mold" a graph into different configurations which can a) better represent her mental model of the data, and b) create views that better communicate the connectivity of one or two target nodes. Figure 4(b) is reached from the state in Figure 4(a) in a simple mouse drag



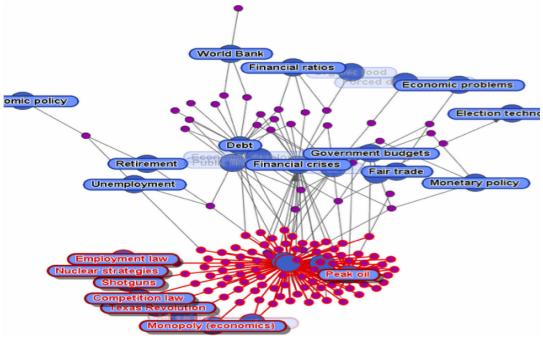
(a) A force directed layout of New York Times articles showing labeled topics. Articles are connected to topics if their LDA association is above a threshold.



(b) An interpolated layout based on results for the search query "oil". These nodes have been dragged to the right and are highlighted in red.



(c) An interpolated layout performing a pairwise comparison. From the previous view, the node for "Nuclear Strategies" has been dragged to the bottom of the screen to achieve this layout.



(d) An interpolated layout based on dragging a cluster of nodes selected from the seed node "Nuclear Strategies" using EvoCut clustering.

Figure 4: Example of interpolation-based layouts.

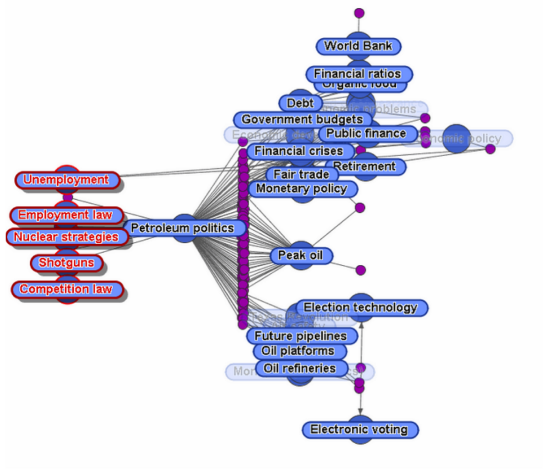


Figure 5: Example workflow in LinkScope, showing an interpolated layout of filtered NSF data.

with our method. Now the graph is arranged in a tree-like fashion, pivoted around the three nodes containing the search keyword "oil";

4.2 Coupling Interpolation with Local Clustering Structure

While the interpolation method can provide insight in some cases and is highly scalable, it only works well on trees and mesh-like graphs where a few click-and-drag would deform the graph in a natural looking way. For highly connected graphs with small diameters, the distance-based interpolation would generate less meaningful layout as it will destroy the nice local layout structures generated by static layout algorithms. To address this issue, we have combined the simple interpolation method with a number of different clustering approaches to allow selecting and dragging a entire cluster while maintaining the internal layout generated by the force directed layout algorithms.

The general idea is described as follows: the analyst select a local cluster based on certain criteria, the central node inside the cluster is used as the pivoting node and moved according to input mouse gesture. The weights of other nodes are computed using the interpolation method. However, for nodes inside the cluster, their weights will be constrained to 1 so they will moved exactly the same as the pivoting node. The weights of the rest of the nodes are decayed further so they would be moved further less to achieve better 'separation'. We also allow 'local refinement' by running different layout algorithms inside the cluster. The relative position of this cluster to the rest of the graph is registered. So if the analyst select another local cluster and choose 'Freeze Clustered Nodes', he/she can do the same interaction without affecting previous untangled structures. Thus the analyst can easily examine the relationship between different clusters and see how they are directly or indirectly linked together.

The key point here is how to effectively select meaningful local clusters. We have explored four possible ways and here we discuss the 'pros and cons' for each of them.

Simple neighborhood expansion The first thing we tried is the neighborhood expansion based on hop distance. Usually a one-ring or two-rings neighborhood can gave a basic idea of what's going on around a node. However, sometimes a neighbor node is better off to be included in another cluster and it will be difficult to make this kind of decision without looking further away in the graph.

Geometric proximity based selection The second variant on the interpolation technique focuses on geometric proximity as opposed to graph distance for selection of the cluster of nodes to be dragged. The idea behind this is that nodes which are similar to each other might not be directly connected. (e.g. in NSF data, nodes are indirectly connected based on similar attribute values). Force directed layout algorithms can do a good job of place these nodes in close proximity to each other. Figure 5 shows the result with a single drag from the original layout in our news article example. Figure 7 shows a second example: The graph shows a filtered set of NSF grant with many different node clusters. The set is filtered to show only grants associated with CA and MA and two other states. The larger yellow nodes represent the states, and the smaller grant nodes are only connected to each other via state nodes. In such cases a topology-based clustering algorithm would fail to find the clusters that are visible in the layout, since there are no direct interconnecting edges between them. We employ a HashGrid method to isolate proximity clusters where edges are not necessarily present. For the simple example shown in Figure 7(c), dragging the cluster that lies between CA and MA is analogous to executing a SQL query with a where clause on the STATE field. The advantage of the visualization is that an analyst can perform a relative analysis of many such queries at a quick glance.

'EvoCut' local clustering algorithm To find a local cluster around a target node in a smarter way, we implemented the EvoCut local clustering algorithm designed by Andersen et al. [8]. A detailed description and complexity analysis can be fount at the original paper. The algorithm tries to find a local partitioning of the graph around a target node by simulating the volumed-biased evolving set process, which is a Markov chain on sets of vertices. It stops either when the desired conductance is achieved or the cost has exceeded a certain value. To our knowledge, EvoCut is by far the fastest local cut algorithm which offers a good balancing guarantee. However, as the process is based on a random work, the size of the cluster and the member nodes may vary significantly each time the analyst probes the same node.

Content-based Selection A logical application is to examine interpolation on node selections based on content. Figure 4(c) shows a further application of the method for the pairwise comparison of targeted nodes. In this case, the analyst is interested in learning about the topical relations between "Oil" and a one of the other labeled topics in the set: "Nuclear Strategies" (NS). The analyst clicks on the representative node in Figure 4(b) and drags it towards the bottom of the view, causing the graph to smoothly transition to the state in Figure 4(c) which is a clustered representation of all graph nodes based on their relative distance from nodes containing the keyword "oil" and the topic node representing "Nuclear Strategies". Interestingly, in this view, the topic node that lies directly between the 'oil' cluster and the "Nuclear Strategies" node is labeled "Petroleum Politics", which seems to make sense semantically. Note that topic labels were assigned manually by a third party expert based on a study of the term lists for each topic.

4.3 Dynamic Graph Modification

So far we have discussed the interactive components in LinkScope that manipulate the graph, but do not modify the link structure of the graph. The following tools in the system modify the graph in some way thereby altering the complexity and in turn, the clarity of views produced by the tools in the previous section. Earlier, a discussion of various graph building functionality was provided. This was in the context of a pre-processing step applied before visualization in an analyst's workflow. However, LinkScope also supports these earlier graph building steps (e.g: edge building and filtering) at visualization time, operating only on currently selected nodes in the visualization. In addition, the system supports a number of other graph altering functions though interaction:

Add/Delete Simple addition and deletion of nodes and edges is supported through mouse selections and keyboard shortcuts.

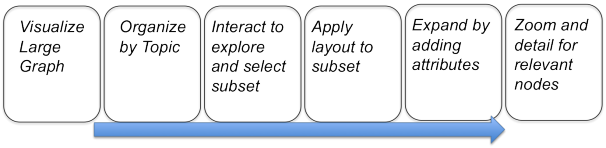
Dynamic Filtering Application of any of the earlier graph building steps based on currently selected input from the visualization.

interactive topic-modeling Gretarsson's iterative topic modeling algorithm [17] is included in the analysis toolkit, allowing a set of topic nodes to be iteratively refined by instantiating the modeling algorithm on a target node-set selected through the interface. This functionality addresses a limitation discussed by Liu et al. [24], in that graded or probabilistic connectivity can be applied between nodes in a dynamic and iterative manner.

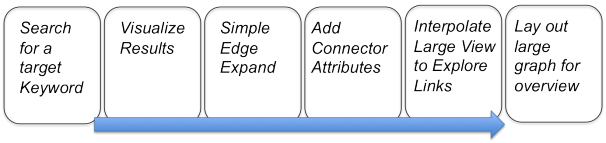
Collapse/Merge/Expand A number of expand, merge and collapse algorithms are available in the toolkit. For example. an analyst has the ability to collapse all documents authored by an individual into a single representative node.

4.4 Example Workflows

The classification of data analysis types outlined in Figure 1 is independent of workflow. Importantly, the toolkit supports a diverse scope of visual analytics workflows. We now discuss a representative example for both top-down and bottom-up analyses.



(a) Example top-down workflow



(b) Example bottom-up workflow

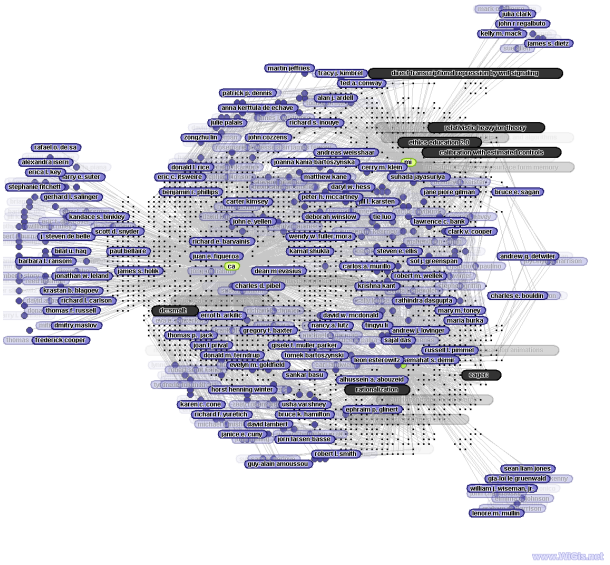
Figure 6: Two example workflows in the LinkScope toolkit. In addition to the linear flow, workflows can be applied iteratively for more complicated analytical tasks.

An analyst can begin with a fine grained starting point such as a targeted search query, a particular individual, grant, topic or institution for example, and perform a variety of expansion functions over the network to arrive at a a broader informative visualization related to the initial seed. Figure 6(b) shows an example of a bottom-up workflow using the available tools. Importantly, LinkScope supports iteration over the steps in the workflow, enabling an analyst to handle complex tasks where necessary. A provenance view for workflows is also supported, as illustrated in Figure 3. Figure 6(a) shows an example of a top-down workflow where the analyst begins with a large graph visualization and invokes various filtering and search tools to arrive at content descriptions for relevant or interesting nodes.

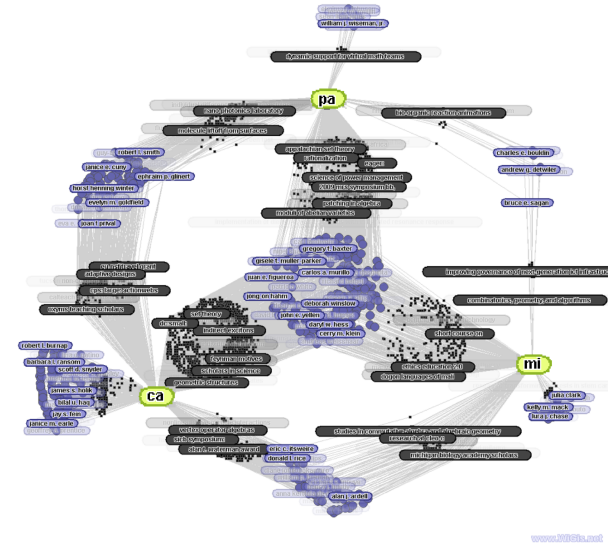
5 Case Studies

In Section's 3 and 4 we have shown some of the interpolation methods working at a high level on on NSF data and on a corpus of New York Times news articles. Now we presnt two more detailed query driven use cases, starting with our NSF data and then on a corpus of incident reports about the conflict in Northern Ireland from the CAIN database [1].

(a) NSF Grants by State: FR Layout



(b) NSF Grants by State: FM3 Layout



(c) NSF Grants by State: Interpolation-based Layout

Figure 7: Multiple layouts of 3905 NSF awarded grants, with 6914 edge connections, for the states of PA, MI and CA from 2008 to 2010.

5.1 Analyzing NSF Grants

To illustrate some of the synergies between dynamic processing of structured data and interactive interpolation of resulting visual graphs in LinkScope, a corpus of 16,561 awarded NSF grants were loaded into the system for analysis. We begin with a typical question that a high-ranking official might ask:

- *"For a given set of US states, who are the most active NSF program managers, and what grants did they fund?"*

The first step in our analysis is to apply a filter for grants that are connected with our target states. In this example, CA, PA and MI were chosen randomly. Following this, an edge builder was applied to produce a graph of all grants awarded in the target states. Next, the PROGRAM_MANAGER property was selected and representative nodes were added to the graph. Again, an edge builder linked them to the appropriate grant nodes. An edge filter ensured that only those PMs related to the selected states were included.

Several attempts were made to visualize the resulting graph of 3905 nodes and 6914 edges. Figure 7(a) shows an attempt to lay out the graph with a simple FR algorithm, and similarly Figure 7(b) shows a layout of the data using FM3. Both algorithms produce cluttered looking layouts, with only a handful of clusters visible in the FM3 case. Next our interpolation method was applied to the graph. Three mouse drags were used on state nodes to produce the layout shown in Figure 7(c). The resulting interpolated layout revealed a far more interesting structure than the other layout mechanisms. In this example, program manager nodes are larger and blue, grants are black and state nodes are yellow and largest. A quick glance at the visualization provided enough insight to answer the probe question above. A proximity-based selection was made one of the discovered sub-clusters –in this case, PMs who awarded grants to CA and MI only (the blue cluster at the bottom of Figure 7(c)), and an interpolation was performed over the cluster. Figure 9 shows a list of these PM names on the right side, and arranges the remainder of the graph (including the particular grants they awarded) relative to that list. Additional general information about distribution of grants by PMs across the three states also becomes clear from the visualization. For example:

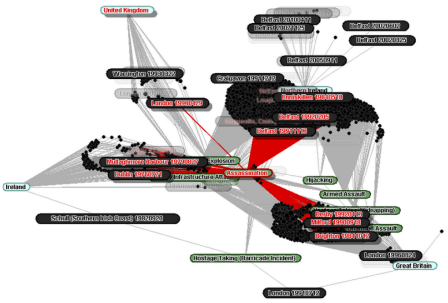
- The PM who awarded the most grants was Jong On Ham, who awarded 493 in total, 46 of which went to CA, 2 to MI and 3 to PA. This information was highlighted by selecting the most connected components using a Dk technique [25] from the LinkScope statistics viewer.
- The total number of PMs across the three states was 443, and they awarded 3459 grants (gleaned from the statistics view panel on proximity-cluster selection). The majority went to CA (bottom left cluster in Figure 7(c) and less than a quarter of this amount went to MI (bottom right cluster).
- Most PMs awarded grants to all three states (central cluster), while very few awarded grants to just one state, with the possible exception of CA (bottom left cluster).

5.2 Terrorism Analysis

To further demonstrate that our interaction methods can provide insight on diverse data, The CAIN dataset [1] –a corpus of police incident reports related to the conflict in Northern Ireland, was analyzed using LinkScope. The raw data consisted of incident reports which were largely text based, with associated meta-data to classify the incidents. For example, reports contained specific fields for country; attack type; success(binary); target type (military, civ. etc.) and property loss (binary). In this example we begin with two sample queries *What was the relative distribution of incidents by country?*, and *What was the relative distribution of assassination attempts*

(a) Layout with only country attribute applied. Clusters are very clear.

(b) FM3 layout after addition of "ATTACK_TYPE" attribute nodes. (9000+ edges)



(c) View after interpolation method applied. Country clusters can be visually resolved. ASSASSINATION node is highlighted, showing that the vast majority of assassination attempts happened in Britain

Figure 8: Three views of the CAIN dataset of 4776 incident reports about the "troubles" in Northern Ireland and Britian from 1970 to 1990.

across the countries?". Figure 8 shows three views of the data in LinkScope used to explore both queries. For the first, every incident was treated as a node on the LinkScope graph, and the COUNTRY property was selected from the edge builder drop-down list in the right panel. Since each incident took place in only one country, a simple force layout reveals the relative distribution of incidents by cluster size in Figure 8(a). The second task is a little more difficult: even though we are applying projection and linking incidents by attribute nodes (as opposed to direct links), in some cases a highly complex graph is returned, even with a small number of attributes. An edge builder was applied and the ATTACK_TYPE property was added to the graph. An FM3 layout of the resulting graph (4776 nodes, 9700 edges) is shown in Figure 8(b). In this view, it is very difficult to visually perceive the country clusters. At this point, an attribute type filter could be applied to only show attack types with the 'assassination" property, which produces a simple graph that answers the query. In Figure 8(c) we show the interpolated graph view after a small number of drags using our cluster based interpolation. The node representing ATTACK_TYPE="assassination" is highlighted in red. Despite the highly complex nature of the data, it is still possible to visually discern the country clusters, and the red edges to each cluster highlight the distribution of assassination attempts in the CAIN data.

6 Related Work

In this section, we discuss the state of the art in visual analytics. Our approaches are compared to others that attempt to combine interactive techniques with dynamic probing of additional attributes and values that can modify the structure and semantics of a graph.

A diverse collection of tools and applications for graph visualization and interaction are available either commercially or as open-source projects. For example, [3, 17, 16, 4, 5, 7, 9] describe tools with a variety of approaches to visualization and feature analysis in graphs. Meta toolkits have been developed to integrate tools from some of these systems, for example the Fekete et al.'s [15] Obvious system. Some tools focus on visual exploration of data from diverse or heterogeneous sources such as ManyNets [16], while systems like Liu's Ploceus [24] have rich functionality for extraction of data from relational tables. Like Ploceus, Our LinkScope system does focus on the ability of an analyst to construct a diverse scope graphs by manipulating attributes from underlying tabular data, however in contrast to Ploceus, LinkScope places more emphasis on an analyst's ability to interactively explore the newly constructed graph though techniques such as interpolated node-specific layout for example.

The importance of ingesting tabular or relational data into a visual representation has been noted in the literature and a number of good tools exist for this purpose. For example [12, 26, 21, 29, 18, 19, 34, 28, 32, 37, 38] all apply interactive preprocessing steps for the dynamic construction of the visualized data set. TableLens [30] Tableau [6] and TouchGraph [2] employ different visual mechanisms for exploring tabular data, including line charts, bar charts and scatter plots. However they do not particularly focus on the analytical process of adding meta data nodes to the graph, as we describe in Figure 2 (e) and (f) earlier. Bostandjiev [12] decribe a user-initiated but largely automated approach to addition of metadata nodes based on underlying structured data, by augmenting a visual search tool on Wikipedia with underlying metadata queried from structured RDF. Moreover, [12] explain how the generated visualization can be used to verify metadata correctness and elicit new metadata from information analysts through interactions. The Jigsaw system [32] semantically relates documents based on lexical comparison of semantic entities within documents. Topic modeling approaches, discussed below are a further step in this direction but abstract the semantic relation to a term set or 'topic' [11].

LinkScope was designed to facilitate interactive graph modification though addition of metadata nodes from underlying tabular or relational data. Given that attribute values can be anything from simple text to nominal, ordinal or discretized numerical values, the ability to sort, organize and compare attribute values is an important design consideration. May [26] explore this issue in the context of generating an interactive visualization. Their SmartStripe method [26] allows a user to "step in to the feature subset selection process" to investigate relations and interdependencies between attributes and attribute values. Currently, LinkScope enables an analyst to perform such comparisons based on interpolated interactions with a node link graph structure. A logical next step for the system is to facilitate more detailed comparative analysis of attributes and values to allow for more fine grained querying. This challenge can be approached either directly, through informed graph manipulation with visual controls and queues, or in an associated tabular view that supports ranking, binning, pivoting and other comparative analyses for attribute values.

In addition to systems that support visual analysis over structured metadata, another relevant class of algorithms focuses on more ephemeral or probabilistic mechanisms for drawing inferences between data nodes in a graph. These methods are generally best suited for analysis of free text. Techniques such as latent semantic analysis, matrix decomposition and factorization methods such as Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and neurocomputation methods such as self-organizing maps have been popular for producing visual representations of large bodies of text. For example, Luminoso [31] and GGobi [33] are text visualization systems that both employ SVD. Computations can be memory intensive for SVD and the resulting topics are not easily interpretable. Other multivariate analysis techniques are also popular in analysis of large text collections. For example Van Ham's PhraseNets [36] supports search for user-provided bi-grams (word pairs), which are then used to drive graph visualizations of large texts. Tools such as PhraseNets are exploratory and the result graphs produced (e.g.: a network of bi-grams) can be easily explored using interactive techniques such as LinkScope, most notably when a seed query returns a highly connected or very large graph. More recently, topic modeling [11, 10] has been used to infer associations between documents, and many tools have been developed [14, 13] to harness topic associations for visual analysis. For example, Gretarsson et al. describe an approach to searching a large corpus of documents based on topic association, where an interactive graph of documents and topics is visualized and refined through iterative application of an LDA [11] algorithm on a selected subset of visualized documents.

7 Implementation

LinkScope is a web based application that uses core Java libraries from [17] to generate graph visualizations. The system employs a client server architecture that captures mouse interaction in the browser and streams them to a remote server. The server component continually computes visual renderings based on incoming input and streams bitmap representations of the current graph view back to the client browser. The advantage of this architecture is that the system can scale well without relying on potentially unknown client-side processing resources.

8 Discussion and Conclusion

A key challenge in visual analytics is to deal with the uncertainty about properties such as size and connectivity of a result set as new search attributes and values are considered by the analyst. To address this challenge, we have introduced LinkScope, a toolkit for performing visual interactive analysis of structured or free text data, primarily though node-link data representations. The system supports three main tool types: 1) graph building tools and filters based on underlying tabular data, 2) visual components such as automated layout algorithms and parallel statistical representations,

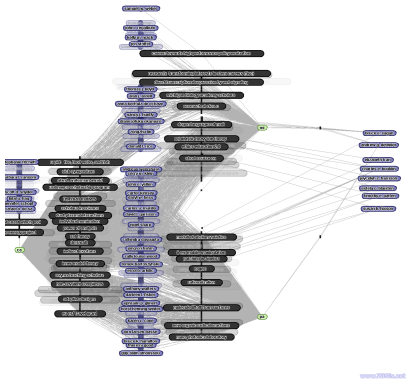


Figure 9: Interpolated view using proximity-based cluster selection. The view highlights a list of the program managers who awarded grants to both California and Michigan from 2008 to 2010. The remainder of the graph is structured relative to this node list.

and 3) novel interactive tools for graph manipulation and exploration. Our research focused heavily on the interaction components, contributing several variants of a novel interpolation layout method based on mouse gestures. We believe that these method are useful for interactive analysis of node-link graphs, particularly when then can be coupled with dynamic attribute extraction from tabular data, and work best when attribute values are represented on the graph as connected nodes. The key innovation of the interpolation techniques is the application of methods traditionally reserved for automated graph layout and clustering to the task of interactive analysis. To evaluate our toolkit and particularly the interactive methods, we presented three analysis scenarios which required multiple graphs to built from tabular data, then analyzed interactively. These data included a corpus of New York Times text-only news articles, arranged by topic, a corpus of 16K awarded NSF grants between 2008 and 2010 with rich metadata and a database of incident reports related to the conflict in Northern Ireland from 1970. In each case our scenarios have shown how interactive interpolation methods can answer targeted analyst questions while revealing interesting related insights which may not have otherwise been considered by the analyst in the initial probe questions.

REFERENCES

[1] Cain project: Conflict archive on the internet.

[2] Touchgraph navigator. Available at www.touchgraph.com.

[3] Gephi : An open source software for exploring and manipulating networks, aug 2010.

[4] Netminer - social network analysis software, aug 2010.

[5] Pajek - program for large network analysis, aug 2010.

[6] Tableau software, aug 2010.

[7] E. Adar. Guess : A language and interface for graph exploration. *Engineering*, (Figure 1):791–800, 2006.

[8] R. Andersen and Y. Peres. Finding sparse cuts locally using evolving sets. *Proceedings of the 41st annual ACM symposium on Theory of computing*, page 20, 2008.

[9] D. Auber. Tulip-a huge graph visualization framework. *Graph Drawing Software*, pages 105–126, 2003.

[10] D. M. Blei and J. D. Lafferty. Topic models. *Text Mining Theory and Applications*, 3(2):113–120, 2009.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[12] S. Bostandjiev, J. O'Donovan, C. Hall, B. Gretarsson, and T. Höllerer. Wigipedia: A tool for improving structured data in wikipedia. In *ICSC*, pages 328–335, 2011.

[13] A. J. Cowell, M. L. Gregory, J. Bruce, J. Haack, D. Love, S. Rose, and A. H. Andrew. Understanding the dynamics of collaborative multi-party discourse. *Information Visualization*, 5(4):250–259, 2006.

[14] J. O. Donovan, A. Asuncion, and D. Newman. Topicnets : Visual analysis of large text corpora with topic modeling. *ACM Transactions on*, V(5952):1–26, 2011.

[15] J.-D. Fekete, P.-L. Hemery, T. Baudel, and J. Wood. Obvious: A meta-toolkit to encapsulate information visualization toolkits - one toolkit to bind them all. In *IEEE VAST*, pages 91–100, 2011.

[16] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. Manynets : An interface for multiple network analysis and visualization. *Challenge*, pages 213–222, 2010.

[17] B. Gretarsson, S. Bostandjiev, J. O'Donovan, and T. Höllerer. Wigis: A scalable framework for web-based interactive graph visualizations. In *GD'09: Proc. of the Intl Symposium on Graph Drawing*, 2009.

[18] D. Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[19] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, 2003.

[20] S. Hachul and M. Jnger. Large-graph layout with the fast multipole multilevel method. *Spring*, V(December):1–27, 2005.

[21] J. Heer, S. K. Card, and J. A. Landay. *Prefuse: a toolkit for interactive information visualization*, volume 05pp, pages 421–430. ACM, 2005.

[22] D. Keim, G. Andrienko, J.-D. Fekete, C. Grg, J. Kohlhammer, and G. Melanon. *Visual Analytics: Definition, Process, and Challenges*, volume 4950, pages 154–175. Springer, 2008.

[23] Y. Koren and A. Civril. The binary stress model for graph drawing. *Graph Drawing*, 2(1):193–205, 2009.

[24] Z. Liu. Network-based visual analysis of tabular data. *October*, 3(c):39–48, 2011.

[25] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic topology analysis and generation using degree correlations. *Proceedings of the 2006 conference on Applications technologies architectures and protocols for computer communications SIGCOMM 06*, 36(4):135–146, 2006.

[26] T. May, A. Bannach, J. Davey, and T. Ruppert. Guiding feature subset selection with an interactive visualization. *Most*, pages 109–118, 2011.

[27] M. Mendoza, B. Poblete, and C. Castillo. *Twitter under crisis: can we trust what we RT?* ACM Press, 2010.

[28] J. O'Donovan, B. Gretarsson, S. Bostandjiev, T. Höllerer, and B. Smyth. A visual interface for social information filtering. In *CSE (4)*, 2009.

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

[30] R. Rao. Tablelens : A clear window for viewing multivariate data. *America*, pages 1–5, 2006.

[31] R. Speer, C. Havasi, N. Treadway, and H. Lieberman. Finding your way in a multi-dimensional semantic space with luminoso. *Media*, pages 385–388, 2010.

[32] J. Stasko, C. Grg, and R. Spence. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[33] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444, 2003.

[34] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. *Learning*, pages 1283–1292, 2009.

[35] P. Trethewey and T. Höllerer. Interactive manipulation of large graph layouts. Tech report, Dept of Computer Science, UCSB, 2009.

[36] F. Van Ham, M. Wattenberg, and F. B. Vigas. Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1169–1176, 2009.

[37] M. Wattenberg. Visual exploration of multivariate graphs. *Proceedings of the SIGCHI conference on Human Factors in computing systems CHI 06*, pages(4):811, 2006.

[38] C. Weaver. Multidimensional data dissection using attribute relation-

ship graphs. *October*, pages 75–82, 2010.