An Application of Edge Bundling Techniques to the Visualization of Media Analysis Results

Wolfgang Kienreich, Christin Seifert
Know-Center, Competence Center for Knowledge-Based Applications and Systems
Graz, Austria
{*wkien,cseifert*}@*know-center.at*

***Abstract*—The advent of consumer-generated and social media has led to a continuous expansion and diversification of the media landscape. Media consumers frequently find themselves assuming the role of media analysts in order to satisfy personal information needs. We propose to employ Knowledge Visualization methods in support of complex media analysis tasks. In this paper, we describe an approach which depicts semantic relationships between key political actors using node-link diagrams. Our contribution comprises a force-directed edge bundling algorithm which accounts for semantic properties of edges, a technical evaluation of the algorithm and a report on a real-world application of the approach. The resulting visualization fosters the identification of high-level edge patterns which indicate strong semantic relationships. It has been published by the Austrian Press Agency APA in 2009.**

***Keywords*-knowledge visualization; media analysis; edge bundling; layout quality evaluation**

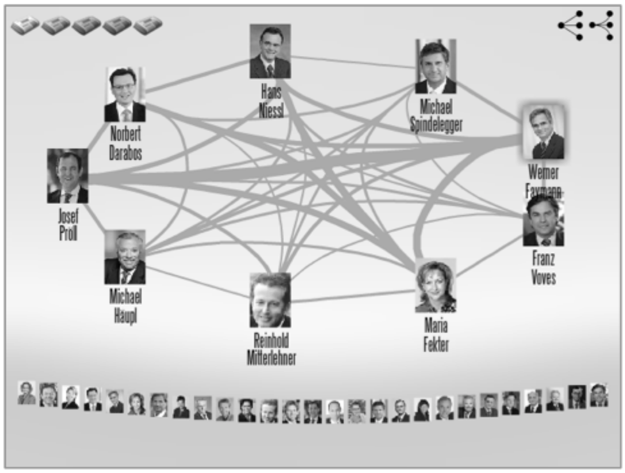


Figure 1. APA MediaConnect Visualization

I. Introduction

The personal digital universe of users around the world has been rapidly evolving over the past decade. The ongoing transformation of the media domain has been a major cause of this evolution: The advent of consumer-generated and social media has significantly increased the quantity, topicality and complexity of the information available through media sources. For instance, commercial and consumer-generated media have published an unprecedented million articles per week covering the US Presidential elections 2008. During the infamous 2008 Mumbai Bombing, the micro-blogging service Twitter provided firsthand information on the terrorist attacks, far in advance of any regular news coverage. The most current images of the event were available on the photo sharing service Flickr. Media consumers have traditionally satisfied personal information needs using a small number of media products. In particular, the information required to participate in public discourse (the communication process enabling collaborative opinion formation and decision making) could be obtained from a limited number of news sources and a single compendium of general knowledge. This approach is no longer sufficient in the light of the outlined expansion and diversification of the media landscape. Nowadays, media consumers have to observe, evaluate and compare a large number of media sources to ensure completeness and correctness of obtained information. They find themselves assuming the role of media analysts in order to satisfy personal information needs. Media providers have recognized this issue and started to address it by providing media analysis services to the general public. Many traditional media analysis techniques, as for example attention tracking, sentiment detection and co-occurrence analysis, can be performed automatically and unsupervised. They are thus suited for deployment in scenarios involving massive numbers of users. However, the results of such techniques are usually of an abstract, numerical nature and cannot easily be interpreted by laypersons. It is therefore essential to devise visual interfaces which present media analysis results in comprehensible form and enable the general public to benefit from the ongoing expansion and diversification of the media landscape.

Information Visualization has frequently been applied to news media repositories [1], [2]. However, the resulting visualizations have usually been tailored towards expert analysts. They have also assumed a high level of visual literacy. In contrast, Knowledge Visualization [3] displays a clear focus on the transfer of knowledge. It frequently employs concrete visual elements or metaphors which are comprehensible to a general audience. Therefore, we consider Knowledge

Visualization a promising framework for the visualization of media analysis results for the general public.

In this paper, we report on the design, implementation and evaluation of an application for the interactive visual analysis of knowledge extracted from news media content. The objective of our work was to create a knowledge visualization capable of conveying complex semantic relationships between key political actors to a general audience. To this end, we have integrated edge bundling techniques with the node-link-diagrams commonly used in representing semantic structures (e.g. [4]). The resulting visualization (see Fig. 1) requires limited visual literacy for interpretation and fosters the recognition and communication of high-level link patterns. It has been available online in the labs area of the Austria Press Agency APA since Summer 2009 [5].

The remainder of the paper is structured as follows: In section II we describe the knowledge we want to communicate and motivate the usage of node-link diagrams. Section III then reviews layout techniques for these kind of diagrams. Our approach including implementation details and data preprocessing is outlined in section IV. The visualization is described in detail in section V and preliminary evaluation results are presented in section VI. We conclude with a discussion and an outlook in section VII.

II. Motivation

Public political discourse in Austria is dominated by a group of approximately 30 political actors. This group is mainly comprised of the members of the elected national and state governments. In the face of the constantly evolving discourse issues and the rapidly increasing number of media sources, it is next to impossible for consumers to manually keep track of the relations between these actors. However, consumers have a concrete need for this kind of knowledge, in order to fulfill their role as the ultimate decision makers in the political process. The automatic, content-based identification of relations between key political actors therefore constitutes a media analysis problem which is highly relevant for a general audience.

Co-occurrence analysis is an automated, statistical approach to content-based relation identification which could provide a solution to the outlined problem. Co-occurrence analysis identifies occurrences of actors in text documents and constructs an actor-to-actor matrix. The cells of the matrix contain numerical values expressing the relatedness of two actors based on their co-occurrence in text documents. The cells of the matrix may also contain information about the textual terms which accompany actor co-occurrences. Unfortunately, the information yielded by co-occurrence analysis is of an abstract, numeric nature. It is therefore not easily comprehensible for a general audience. Hence, we propose to express co-occurrence analysis results through a knowledge visualization. This visualization should enable users to put co-occurrence information into context and accumulate knowledge about the current state of relations between key political actors, as expressed through news content.

III. Related Work

The information yielded by co-occurrence analysis can be represented as a semantic graph. We describe a simple general notation of semantic graphs in section III-A. Semantic graphs are usually visualized as node-link diagrams. The construction of node-link-diagrams involves node placement, link layout and clutter reduction. We discuss appropriate techniques in section III-B.

A. Semantic Graphs

A general graph $G = (V, E)$ is comprised of a set of nodes V and a set of edges E between nodes. The graph is called a weighted graph, if a numerical value is associated to each node and/or edge. In a semantic graph, nodes represent semantic concepts and links represent semantic relations between concepts. Concepts and relations can be manually modeled by domain experts or automatically extracted from knowledge artifacts. In the scope of this paper, we are interested in semantic graphs which have been extracted from news article repositories. We consider semantic graphs in which concepts correspond to political actors and relations correspond to the co-occurrence of political actors in news articles. In terms of data structure, our semantic graph is a weighted undirected graph comprised of a set of nodes V, a set of edges E, a weighting function w_v for the nodes, a weighting function w_e for the edges, and two similarity functions $f_e : E \times E \rightarrow [0, 1]$ and $f_v : V \times V \rightarrow [0, 1]$. The function f_v represents the similarity of two nodes, i.e. the similarity of the semantic concepts associated to the nodes. Similarly, the function f_e represents the similarity of two edges, i.e. the similarity of their associated semantic relations. The semantic graph is schematically outlined in Fig. 2.

B. Node-Link Diagrams

The placement of nodes (node layout) can be understood as a dimensionality reduction problem if each node is associated with a high-dimensional feature vector. Common approaches to dimensionality reduction include Principle Component Analysis (PCA) [6] for linear reduction and Isomap [7] for nonlinear reduction. However, these methods are computationally intensive and do not consider semantic relations which are not encoded in the feature vectors. Self-organizing maps (SOMs) [8] employ artificial neural networks which are trained to produce a low-dimensional, discretized representation of a high-dimensional input space. However, this method is computationally intensive and the implicit vector quantization is not always desirable. A physics-based layout method is force-directed placement (FDP) [9]. Force-directed placement simulates the entire

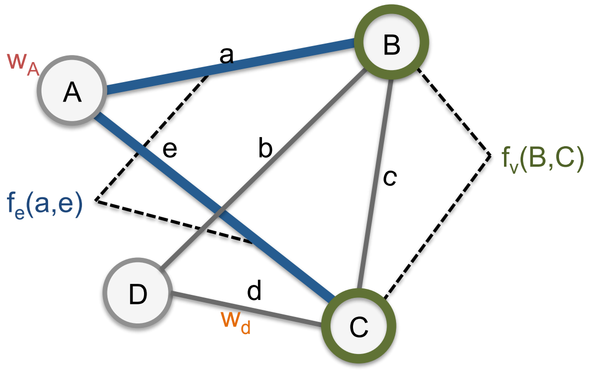


Figure 2. Semantic Graph: Semantic nodes (representing e.g. political actors) are connected by semantic edges (representing e.g. co-occurrences). A weight is associated to semantic nodes and edges. The similarity functions for edges and nodes are not displayed but implicitly used for the layout.

graph as a physical system in which edges act as springs and nodes act as electrically charged particles.

The aim of the link-layout algorithms is the visual ”beauty” of the graph on the one hand and clutter reduction to foster comprehensibility on the other hand. Interactive visualizations of graphs enable the visual representation of a graph to change according to user interactions. This fact has been utilized for clutter reduction, for example by curving graph edges away from the user’s focus of attention while retaining graph node positions [10]. Obviously, these techniques cannot be applied to static visualizations of graphs. Clutter reduction techniques based on edge bundling try to reduce visual complexity by merging edges which satisfy certain criteria. Confluent graph drawing [11] generates planar visual representations of non-planar graphs by merging edges. Unfortunately, it is nontrivial to decide if a given general graph can be visualized using this method. Flow map layouts [12] apply hierarchical binary clustering to a set of nodes, positions and flow data to enable edge merging and routing. Specific bundling and routing methods have been devised for circular node layouts [13]. We have based our work on the force-directed edge bundling algorithm proposed by Holten and van Wijk [14]. This method iteratively subdivides edges and applies forces to subdivision points to guide the bundling process. It can be applied to general graphs in static and dynamic application scenarios and does not require a hierarchy, a control mesh or any other external data structure. However, this force-directed edge bundling works on a purely geometrical basis and cannot express semantic properties of nodes and edges.

IV. Our Approach

Our approach to the visualization of relations between political actors is based on the construction of a node-link diagram using projection methods for node layout and edge-bundling techniques for link layout and clutter reduction. We have implemented this approach in an industry project conducted in cooperation with the Austrian Press Agency APA. The resulting application has been available to the general public under the name MediaConnect since summer 2009 [5]. In this section, we discuss the environment, design and implementation of the application and provide preliminary evaluation results.

A. Data Preprocessing

Data acquisition was based on a list of relevant political actors provided by our application partner. We first collected relevant news articles through search queries for the family names of all actors. The resulting set of articles was preprocessed using a custom information extraction module optimized for German language. This module annotated each article with named entity information, i.e. person names. We could then identify relevant actor occurrences from annotated person names using the initial actor list as a white-list. Based on identified occurrences, we computed the co-occurrence matrix of actors. We also constructed a noun vector space using Part-Of-Speech tags. We employed this vector space in the node and link similarity computations described below. The data structure for our visualization, the semantic graph (see section III-A) is constructed as follows:

nodes: Each political actor identified in the current set of news articles is represented by a node.

node weights: The weight of a node equals the number of news articles containing the corresponding political actor normalized by the total number of news articles.

edges: An edge between two nodes is constructed if the two corresponding political actors co-occur in at least one news article.

edge weights: The number of articles related to an edge equals the number of articles where both political actors associated to the edge’s nodes co-occur. The weight of an edge then is set to this normalized co-occurrence value.

node similarity: The feature vector associated to a node is the combined feature vector of all news articles in which the corresponding political actor occurs. The similarity s of two nodes v_i, v_j is calculated as the cosine similarity of the feature vectors f_i and f_j associated with the nodes.

$$s_{v_i,v_j} = cos(f_i, f_j) = \frac{f_i \cdot f_j}{||f_i|| \cdot ||f_j||} \tag{1}$$

edge similarity: The feature vector associated to an edge is the combined feature vector of all news articles in which its nodes’ corresponding authors co-occur. The similarity of two edges e_i and e_j is calculated as the cosine similarity between their feature vectors f_i and f_j similarly to equation 1.

B. Environment

The MediaConnect application has been implemented as a module within APA Labs, an experimental, web-based platform supporting retrieval and analysis of news articles provided by the archives of the Austrian Press Agency [15]. The APA Labs platform accepts query specifications including query terms, a list of sources and a date range. It returns a result set of news articles including named entities, feature vectors and similarity values. The modular client-server structure APA Labs Framework allows to easily incorporate new visualizations modules.

C. Layout

In general, we compute a two-dimensional layout of nodes using a force-directed placement algorithm based on a lin-log energy model [16]. Our implementation of this algorithm places nodes close to each other if their respective feature vectors are highly similar. The technical evaluation described in section VI was performed on layouts produced by this implementation. However, for the MediaConnect application, a fixed circular layout of nodes was required. We therefore computed a one-dimensional layout using a custom projection algorithm based on inverse feature vector similarity. This algorithm places nodes far from to each other if their respective feature vectors are highly similar. Note, that the largest possible distance on a one-dimensional line is the distance between the line endpoints, but the largest possible distance on the circle is $0.5\times$ the circumference. Therefore, we adapted the projection algorithm to take the circular structure into account. The computed node positions were then projected onto the layout circle.

In the resulting layout (see Fig. 1), edges can be clearly distinguished because strongly connected nodes are placed far from each other. We note that the layout used in the MediaConnect application was based on customer requirements, and that we do not expect this layout to generalize to other scenarios or larger graphs.

For the edge layout we adapted the force-directed edge-bundling algorithm in [14]. In the initial paper the edges are bundled according to four different (geometric) edge compatibility measures C_a, C_d, C_s and C_v:

C_a angle compatibility: edges with similar angles should attract each other to a higher degree, $C_a \in [0, 1]$

C_d distance compatibility: nearby edges should attract each other to a higher degree, $C_f \in [0, 1]$

C_s scale compatibility: edges of similar length should attract each other to a higher degree, $C_s \in [0, 1]$

C_v visibility compatibility: edges which have high overlap when projected to each other should attract each other to a higher degree, $C_v \in [0, 1]$

We introduced a fifth compatibility measure, the semantic compatibility C_e. The semantic compatibility measures the semantic relationship between two edges.

C_e semantic compatibility: edges should only be bundled if they are semantically related, $C_e \in [0, 1]$

We defined the semantic edge compatibility as the semantic similarity defined in equation 1. Semantic edge compatibility could be used as a stand-alone measure or in conjunction with the geometric measures. In our experiments we found out that pure semantic bundling best reflected the semantic relationships between the edges but lead to cluttered graph layout, because it increased the number of edge crossings and the total edge length. In our application we therefore used the combined semantic and geometric edge compatibility to the total compatibility measure C:

$$C = k \cdot \underbrace{C_a \cdot C_d \cdot C_s \cdot C_v}_{\text{geometric}} \cdot \underbrace{C_e}_{\text{semantic}} \tag{2}$$

with k being the global bundling parameter.

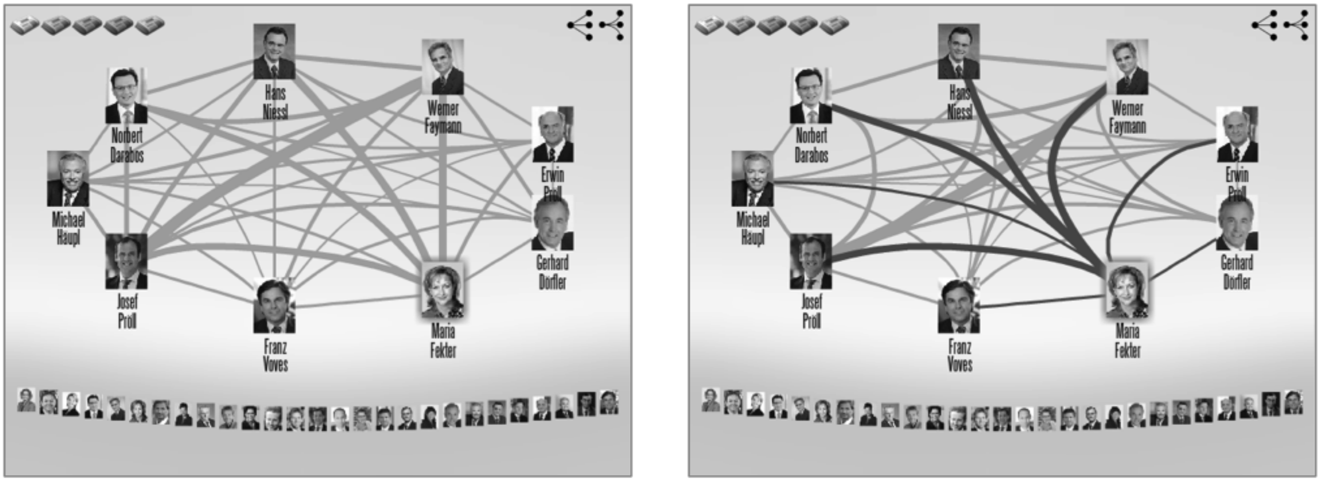
In the current visualizations the weights of the nodes are not visually encoded, the weights of the edges are encoded by line thickness.

V. Design and Interaction

The MediaConnect visualization displays semantic relationships between selected political actors. Users can select an actor from the actor portraits in the gallery on the lower edge of the visualization. Moving the mouse over a portrait in the gallery displays the name and the job description of the actor. Clicking on a portrait selects an actor, and all actors to which this actor has a direct semantic relationship, for analysis.

Selected political actors are represented as labeled portraits arranged in a circular layout above the gallery, in the center of the visualization. The existence of a semantic relationship between two actors is represented by a curve connecting their respective portraits. The width of this curve expresses the strength of the relationship: A wide curve indicates a strong relation. For example, the strongest semantic relationship in Fig. 3(a) is between Werner Faymann and Josef Pröll, Maria Fekter is strongly related to Werner Faymann but weakly related to Erwin Pröll, and Franz Voves does not have a strong relation to any other actor.

By default, MediaConnect visualizes the combined semantic relatedness found in all underlying media sources. Users can select a single media source by choosing from the buttons on the upper left of the visualization to compare source coverage. If a single source is selected, all curves are assigned the color and width corresponding to that source. Users can move the mouse pointer over a single political actor to highlight all curves connecting this actor to other actors. Edge bundling can be activated by clicking the appropriate button on the top right side of the visualization. An animated transition between bundled and straight edge layout is provided to avoid breaking the visual context. Fig. 3(b) displays the effect of edge bundling. A moderate amount of bundling has been applied and, for example, emphasizes the strong relation between Werner Faymann and Josef Pröll. Because bundled edges sometimes make it difficult to identify source and target of a relation, the



(a)

(b)

Figure 3. APA MediaConnect Visualization with (a) direct edges, and (b) bundled edges

highlighting of all outgoing edges is an important feature in the bundled case.

VI. Evaluation

We have conducted a technical evaluation of our approach by computing a set of quality indicators for results obtained from real-world and synthetic datasets. We compared straight edge layout, geometric edge bundling, semantic edge bundling and the combined bundling used in the MediaConnect application. We measured the number of edge crossings, the total edge length, the pixel coverage and the pixel overdraw and obtained significant results by computing paired T-tests. We also measured the correlation between the geometric and semantic distance of edge segments to evaluate the validity of the semantic edge compatibility measure. We found that all forms of edge bundling significantly increase edge length and significantly decrease the number of edge crossings in comparison to the baseline straight edge layout. We also found that the semantic bundling produces a very high correlation between geometric and semantic distance. The combined bundling produces a slightly lower, but still high and significant, correlation.

The results of the technical evaluation validate our approach: Apart from the baseline straight layout, the combined geometric and semantic bundling employed in the MediaConnect application performs best for the number of edge crossings, the total edge length and the pixel overdraw, while providing a high correlation between semantic and geometric distance. We have also conducted an informal, qualitative user evaluation. The Austrian Press Agency provided a small group (five persons) of domain experts, which were familiar with various media analysis tools. This group was given a minimum amount of training and explanation. We then encouraged the group to use the MediaConnect application for some days and collected and integrated written user feedback. The most frequently raised demands were for labeled edges and for control over the degree of bundling applied in the visualization. The most commonly named application scenario was query expansion along the interpersonal relations displayed in the visualization.

VII. Conclusion and Future Work

In this paper, we have investigated how semantic relationships between political actors can be visualized using node-link diagrams. We have described a force-directed edge bundling approach to the reduction of edge cluttering which respects semantic relatedness between links. A technical evaluation has shown that this approach reduces the number of edge crossings and increases the correlation of link geometry and link semantics. We have integrated the components of our approach into a real-world scenario based on news articles provided by the Austrian Press Agency APA. The resulting visualization is available to the general public, and has met the approval of domain experts in informal, qualitative evaluations. We expect our approach to foster the recognition of high-level patterns in the relations between political actors. Technical evaluation results provide some indicators that this expectation could be met. However, formal experiments will be required to verify our claims. Therefore, we will focus future work on the evaluation of our approach.

Acknowledgment

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

REFERENCES

[1] E. Rennison, “Galaxy of news: an approach to visualizing and understanding expansive news landscapes,” in *UIST ’94: Proceedings of the 7th annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM, 1994, pp. 3–12.

[2] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, “The InfoSky Visual Explorer: Exploiting hierarchical structure and document similarities,” *Information Visualization*, vol. 1, no. 3–4, pp. 166–181, Dec 2002.

[3] S.-O. Tergan and T. Keller, Eds., *Knowledge and Information Visualisation – Searching for Synergies*, ser. Lecture Notes in Computer Science. Germany: Springer, Jun 2005, vol. 3426.

[4] J. Shen, L. Li, T. G. Dietterich, and J. L. Herlocker, “A hybrid learning system for recognizing user tasks from desktop activities and email messages,” in *Proceedings International Conference on Intelligent User Interfaces (IUI)*. New York, NY, USA: ACM, 2006, pp. 86–92.

[5] “APA MediaConnect,” Web Application. [Online]. Available: http://labs1.apa.at/ApaLabs/links.jsp

[6] I. Jolliffe, *Principal Component Analysis (2nd ed)*. Springer, 2002.

[7] J. B. Tenenbaum, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000. [Online]. Available: http://dx.doi.org/10.1126/science.290.5500.2319

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, “Self organization of a massive document collection,” *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 574–585, May 2000.

[9] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software - Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, November 1991. [Online]. Available: citeseer.ist.psu.edu/fruchterman91graph.html

[10] N. Wong and S. Carpendale, “Supporting interactive graph exploration using edge plucking,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 6495, Jan. 2007.

[11] M. Dickerson, D. Eppstein, M. T. Goodrich, and J. Y. Meng, “Confluent drawings: Visualizing non-planar diagrams in a planar way,” *J. Graph Algorithms and Applications*, 2005.

[12] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan, “Flow map layout,” in *IEEE Symposium on Information Visualization (InfoVis).*, 2005, pp. 219–224.

[13] E. R. Gansner and Y. Koren, “Improved circular layouts,” in *Proceedings International Symposium on Graph Drawing (GD)*, ser. Lecture Notes in Computer Science, M. Kaufmann and D. Wagner, Eds. Springer, 2007, pp. 386–398.

[14] D. Holten and J. J. van Wijk, “Force-directed edge bundling for graph visualization,” *Eurographics/IEEE-VGTC Symposium on Visualization (Eurovis)*, pp. 983–990, 2009.

[15] E. Lex, C. Seifert, W. Kienreich, and M. Granitzer, “A generic framework for visualizing the news article domain and its application to real-world data,” *Journal of Digital Information Management (JDIM)*, vol. 6, no. 6, Dec 2008. [Online]. Available: http://www.dirf.org/jdim/v6i6.asp

[16] A. Noack, “Energy models for graph clustering.” *J. Graph Algorithms Appl.*, vol. 11, no. 2, pp. 453–480, 2007.