Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations

Stef van den Elzen and Jarke J. van Wijk

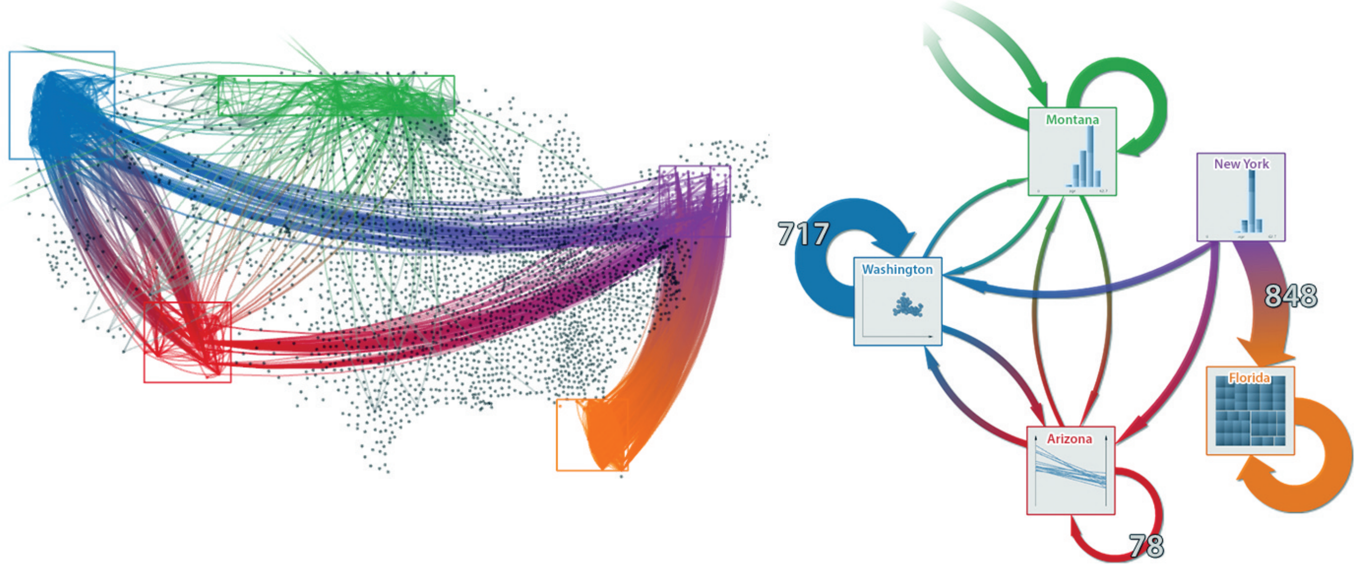


Fig. 1. Multivariate network exploration using selections of interest, detail view (left) and high-level infographic-style overview (right).

Abstract—Network data is ubiquitous; e-mail traffic between persons, telecommunication, transport and financial networks are some examples. Often these networks are large and multivariate, besides the topological structure of the network, multivariate data on the nodes and links is available. Currently, exploration and analysis methods are focused on a single aspect; the network topology or the multivariate data. In addition, tools and techniques are highly domain specific and require expert knowledge. We focus on the non-expert user and propose a novel solution for multivariate network exploration and analysis that tightly couples structural and multivariate analysis. In short, we go from Detail to Overview via Selections and Aggregations (DOSA): users are enabled to gain insights through the creation of selections of interest (manually or automatically), and producing high-level, infographic-style overviews simultaneously. Finally, we present example explorations on real-world datasets that demonstrate the effectiveness of our method for the exploration and understanding of multivariate networks where presentation of findings comes for free.

Index Terms—Multivariate Networks, Selections of Interest, Interaction, Direct Manipulation.

1 Introduction

Many real-world phenomena can be modeled as multivariate networks: e-mail traffic between persons within a company, a telecommunication network, money flowing between bank accounts, or physical objects such as airplanes flying from airport to airport or migration of people between cities. The common theme here is the connection (relation, link, edge) between objects (nodes, vertices). The number of nodes and links of real-world data is generally large, in the order of thousands. For these networks often more information on the nodes and links is available. For example, in case of a company e-mail network we know more attributes of the persons (nodes) involved, like age, gender, and job title. We also have more information about the e-mails (links) such as time-sent, header-information, and body text.

The exploration and analysis of large multivariate networks is still a challenge. Current methods are focused on either the structural aspect of the multivariate network, e.g., [46] or the multidimensional data attached to the nodes and links, e.g., [35]. However, we believe the greatest insights are gained from simultaneous exploration, as the two might be correlated or influence each other. For example, we are not only interested in who is e-mailing to whom (structure) or whether females or males are communicating more (multivariate data), but we are more interested in whether females are communicating more with females or more with males and also between which departments and what the distribution over time is (both structure and multivariate data). For this we need to be able to inspect the attributes in context of the underlying network topology. We provide a method that enables users to explore both aspects in a uniform method using selections of interest as central element. In summary, we go from Detail to Overview via Selections and Aggregations, which explains the acronym we selected for our approach: DOSA. And also, a dosa is a spicy Indian wrap, which resonates with our aim to combine existing ingredients into a tasteful result.

- *Stef van den Elzen is with the Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands, and SynerScope BV, Eindhoven, The Netherlands. E-mail: s.j.v.d.elzen@tue.nl.*
- *Jarke J. van Wijk is with the Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands. E-mail: j.j.v.wijk@tue.nl.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Multivariate networks are commonly visualized using node-link diagrams for structural analysis [36]. However, node-link diagrams do not scale to large numbers of nodes and links and users regularly end up with hairball-like visualizations. The multivariate data associated with the nodes and links are encoded using visual variables like color, size, shape or small visualization glyphs [30].

From the hairball-like visualizations no network exploration or analysis is possible and no insights are gained or even worse, false conclusions are drawn due to clutter and overdraw. For the non-expert user, the large network visualizations are overwhelming, confusing and contain too much detail. The casual user just wants a simple (minimalistic) visualization that conveys a clear message about the relation between network structure and multivariate data. We support both expert and casual users by presenting two juxtaposed coupled views; a detail view with all low-level network elements and a high-level infographic-style overview with aggregated components. During exploration in the detail view, the high-level overview is updated automatically. Exploration and analysis is supported by defining selections of interest. Domain experts can still use advanced measures like network distance and centrality in an uncomplicated and uniform manner, while a simplified overview that can, for example, be used for communication to the non-expert user, is generated for free and can be further refined with minimal effort. The casual user is supported with intuitive controls and playful interaction that encourages to explore the network.

In this paper we propose a novel method for multivariate network exploration and analysis. More specifically, our main contributions are:

- a tightly coupled exploration method, enabling users to explore and analyse both network structure and multivariate data associated with the nodes and links simultaneously, using
- intuitive creation and modification of selections of interest, and
- a juxtaposed detail and high-level overview, for
- effortless production of high-level, infographic-style overviews, focusing on the non-expert user.

The paper is organized as follows. First, related work is discussed in Section 2. Next, our approach to multivariate network exploration and analysis is described in Section 3. We describe the two juxtaposed views in Sections 4 and 6, and explain how exploration is facilitated using selections of interest in Section 5. Next, example explorations on real-world data are given in Section 7 and limitations are discussed in Section 8. Finally, conclusions and directions for future work are provided in Section 9.

2 Related work

The most well-known and widely used method to visualize networks is a node-link diagram. Each object is represented by a dot and if there is a connection between two objects a line is drawn in between. Much work is focusing on computing two-dimensional layouts (embeddings) for node-link diagrams that best convey network topology while taking aesthetic criteria into account to improve readability [4]. Multivariate data associated with the nodes and links is commonly depicted using visual variables, such as color, size, and shape of both the nodes and links [7, 16, 27, 30, 31, 36]. Also, glyphs are used to represent the nodes [43] and motif glyphs enable structural insight [12].

As opposed to emphasizing topological properties of the network, multivariate data can be used to compute attribute-based layouts [3, 15], such as the spherical Self-Organizing Maps [47] and JauntyNets [24], to provide more insight in the multivariate data involved. Furthermore, multivariate data can be used to directly define a layout by using a scatterplot for the nodes and superimposing edges onto this, as in the GraphDice system [5]. Readability of node-link diagrams for large networks is challenging due to overlap, overdraw and clutter in general, this is aggravated further by the use of visual variables to convey associated multivariate data.

A broadly used metaphor to prevent clutter in node-link diagrams are lenses practicing focus+context techniques [6, 37]. Lenses are used to enable inspection of dense areas of the network [46] and show more information for nodes of interest by displaying in-situ visualizations [23] or extract subparts of the network for further exploration [21]. Our solution also involves selections of interest, represented by boxes partially based on ideas of lenses.

A method specifically designed for multivariate network exploration and closest to our technique is Semantic Substrates [35]. In Semantic Substrates non-overlapping regions are introduced representing different categorical node attributes. In each region, nodes can be layed-out directly according to the node attribute values or nodes positions can be computed via a force-directed layout algorithm. Edge visibility is controlled via graphical user interface controls to prevent clutter; for each region, visibility of an edge to another (or the same) region can be set. Our selections of interest are similar to the non-overlapping regions of Semantic Substrates, albeit more flexible. Semantic Substrates regions are restricted to a single categorical node attribute and link attributes are not taken into account. We support both *n*-dimensional regions as well as link attributes. Furthermore, link visibility is controlled globally, while we implement a more fine-grained local region control.

PivotGraphs provide an aggregated view on the network by showing two axes with categorical node attributes and positioning the nodes on the grid according to their associated attribute values [44]. This provides abstraction and a means to explore categorical node attributes supported by pivot and roll-up operations, inspired by database OLAP (*online analytical processing* [9]) actions. Unfortunately, due to aggregation, network topology is not preserved, turning structural exploration into a challenge as multiple operations need to be performed for a comprehensive image. Aggregation is also used in the GraphTrail system [11] for multivariate network exploration. Here the focus is mainly on capturing the user interaction and integrating this into a history trail. Familiar charts are shown for the exploration of the multivariate data.

Pretorius and Van Wijk [28] enable multivariate network exploration by treating links as first class citizens. Link labels are placed in sequence top-to-bottom in a rectangular region centered between source and target nodes on both sides. Each node is contained in a hierarchy defined by associated multivariate data rendered as an icicle plot that is positioned on both sides of the edge labels. Next, each node is connected with a line to the according edge label. This is extended to multiple hierarchies in the Parallel Node-Link Bands approach [18]. Users can interactively inspect and query the graph, however, due to the bipartite node layout it is difficult to explore network topology.

The field of multivariate network visualization and interaction is large and we only discussed the most relevant related work. For a more complete overview, we refer to survey papers on the visual analysis of large graphs [19, 41] and a recent book on multivariate network visualization [25].

In summary, current methods are focused either towards structural exploration or multivariate data exploration. No method facilitates both the structural and multivariate analysis in a tightly coupled exploration technique. Also, no system provides users with an easy to understand simplified overview showing both structure and associated multivariate data, except for PivotGraph, but there the low-level details are missing.

3 From detail to overview

Large multivariate network exploration is a challenge due to size and inability to explore node and link attributes in context of the underlying network topology. Furthermore, to non-expert users a low-level visualization showing all individual elements is overwhelming, confusing and provides too much detail. They rather need an aggregated overview showing the most important components. Also, the expert user needs this as a means of communication to stakeholders. In summary, to support this, we need:

- a scalable interactive method to simultaneous explore network structure and associated multivariate data for the nodes and links using direct manipulation, and

- the ability to see both the low-level details and aggregated high-level elements, all using
- familiar metaphors.

To tackle the scalability problem there are two main approaches, top-down [34] and bottom-up [39] exploration. In a top-down approach, exploration starts with an overview of the entire network. From this overview interesting features are identified and the exploration continues with a more narrow focus on sub-structures of the network. This is difficult with a large node-link diagram; due to clutter and overdraw interesting features are hard to discern. Inversely, a bottom-up approach starts with a (predetermined) single node of interest and then continues the exploration to neighboring nodes.

Here we pursue a hybrid approach; we do not limit the exploration to just one node but to a number of selections of interest (see Section 5) that each contain one or more nodes and we simultaneously always show both the low-level detail (see Section 4) and high-level overview (see Section 6). Our novel DOSA exploration process using the described elements is schematically shown in Figure 2.

4 Detail view

To provide an overview of the network, each individual node is shown in the detail view. The position of the nodes is determined by a customizable two-dimensional projection of the network based on node attributes. Note that this is not limited to a scatterplot-like visualization, but can also be a (precomputed) force-directed network layout or more familiar geographic plot as both can be encoded via attributes of the nodes. The axes involved in the projection can be shown or hidden on demand. Animation is used to show the relation between projections upon change and maintain the notion of a unified information space. Users are enabled to freely zoom-and-pan the projection space to navigate and explore.

To prevent clutter, edges are initially not shown. We only show the edges involved in the selections of interest, more on this is described in the next section. The two main approaches to depict edge direction are arrows and color, e.g., [20]. Here, we choose to render edges using quadratic curves in a clock-wise fashion to convey directionality (see Figure 3a). This prevents overdraw of bidirectional edges, avoids clutter because arrow-heads do not have to be drawn and finally, the visual variable color can be used to convey a different attribute, here visual association between different selections of interest.

Alongside the available multivariate data at the nodes, we also compute structural network properties for each node such as degree and centrality measures closeness and betweenness. By changing the projection, exploration can start from an interesting multivariate data property, e.g., cities with a low population and high crime rate, or from an interesting structural node property such as high betweenness, a geographical region, or a combination of these. The creation and interaction with these selections of interest is described in the next section.

5 Selections

In the multivariate network exploration process, users need to be able to focus on subparts of the network and then aggregate these to perform high-level comparison and inspection. For the selection of interesting subparts, the following candidate solutions can be employed:

- *brushing selection*: one set consisting of the current brushed items is highlighted, the rest of the items are treated as background.
- *partitioning*: multiple sets of items, supported by, e.g., brushing with different colors in Xmdv [26, 42] or automatically coloring of items by conditional formatting as in Microsoft Excel.

There are two underlying principles that enable the creation of selections: *painting* and *querying*. For painting the elements are pointed at with the mouse cursor and colored accordingly. For querying, a list of predicates on attributes is specified.

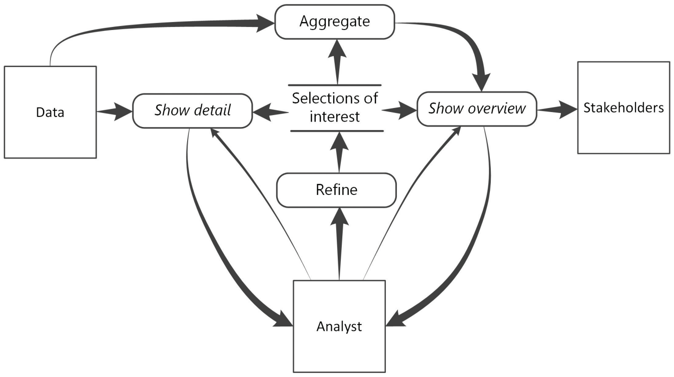


Fig. 2. The DOSA exploration process with selections of interest as central element. The analyst (bottom) refines selections of interest, which influences both the detail and overview visualizations from which insight is gained on a low detail and aggregated high level. Finally, insights can be communicated directly to stakeholders in a simplified, infographic-style visualization conveying a clear message.

We want a method that is both *expressive* and *simple*. However, in general these requirements are conflicting, for example the DataMeadow approach [14] is expressive but complex. We selected to use an approach based on partitioning and (visual) querying, because this provides better support for scanning and exploring the data.

We provide a visual querying mechanism to explore multivariate networks based on a node partitioning. Our solution for node selection is based on the following familiar metaphors:

1. draw boxes,
2. select ranges,
3. order selections (similar to arranging layers in Adobe Photoshop, or arranging objects in Microsoft PowerPoint), all using
4. direct manipulation.

Users can create selections of interest by adding boxes to the current projection in the detail view. The nodes belonging to this selection are the nodes contained in the box, i.e., within the ranges of the two projected attributes. Users can freely reposition the box by dragging in the projection, this dynamically changes the ranges of the selection. Also, the size of the box can be adjusted via standard selection controls, such as drag handles, to directly influence the associated ranges. We support users to quickly set a similar attribute range to all selections. If this attribute option is active then the according range is synchronized over all boxes.

We choose for boxes here over other alternatives such as freeform selection to support intuitive simple interaction: the selections of interest are easy to visualize, and, the boundaries of the box directly translate to ranges for the two projected attributes in the detail view, this enables intuitive and simple manipulation, in both the detail view and the scented widget controls.

Boxes and contained nodes have a (adjustable) color for visual association. Upon changing the projection, the previously defined ranges for a selection of interest are maintained. The position, width and height of the box are adjusted to reflect the current ranges for the projected attributes. Users are enabled to shift their focus to specific boxes while maintaining a context using smooth zoom and pan methods [40].

Each selection is shown in a selection component (see Figure 3d), that has additional operations such as hide and lock, similar to the layer approach in Adobe Photoshop. Furthermore, the color and name of a selection can freely be changed to something semantically meaningful. The selection component provides a box selection mechanism and

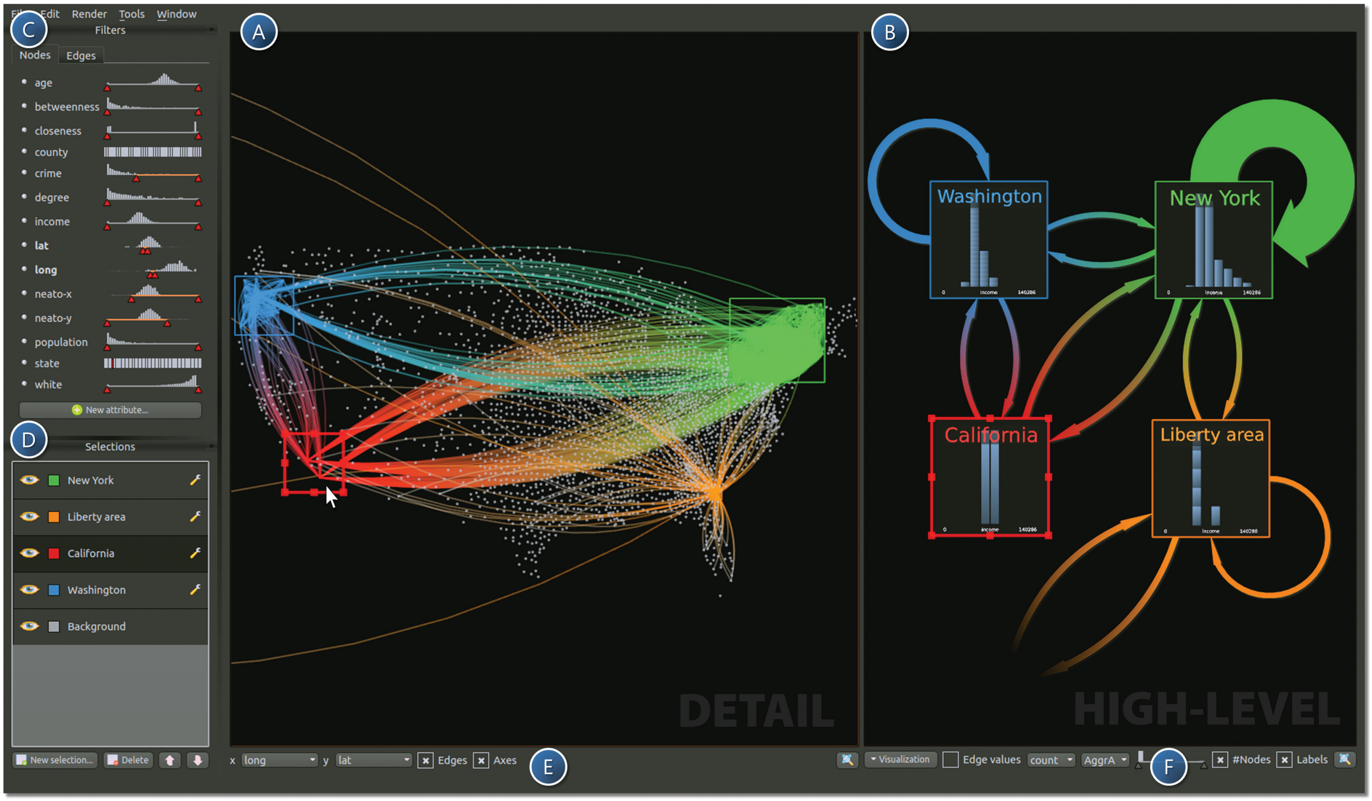


Fig. 3. Graphical user interface of the implemented prototype showing all coherent components: a) Low-level detail view showing a two-dimensional projection of the nodes based on available attributes. The projection and other visual attributes can be set using controls at the bottom (e). Four selections of interest are shown in the detail view, visualized using boxes for direct manipulation. All selections of interest show *between* edges, the green, blue, and, orange selections also show *within* edges. The orange selection additionally shows edges with the *background* selection. b) High-level overview showing aggregations of the selections of interest including associated aggregated edges. For each of the selections an interactive histogram visualization is shown. Visual representation and attribute mapping are configurable to users needs with controls at the bottom (f). c) Attribute component showing all available attributes with according Scented Widgets for the nodes and links in different tab-pages. The Scented Widgets provide information on the distribution of attributes and can be used to directly control the ranges of the multidimensional selections of interest. d) Selection component containing a list of all selections. Selection priority (order) is controlled via drag and drop operations. Additionally, selections can be hidden or locked here.

simultaneously serves to resolve conflicts in the selections; as a result of our partition approach, a node can only belong to a single selection. If there is overlap of the boxes, then the order of the selections is decisive. Selections higher in the list bind stronger. The sequence of the selections can be changed using drag and drop or button controls to enable fast switching of possible box configurations in the selection component. The selection order can also be influenced using a context menu with arrangement controls on the boxes in the detail view, e.g., bring to front, sent to back.

Initially users are provided with a single *background* selection that contains all nodes. This approach is twofold: it provides an overview of the entire network showing dense and sparse areas to start the exploration, and it provides a context to the selections made.

The underlying formal model with technical details on the realization is described in the next section.

5.1 Model

We have a network $G = (N, E)$ with nodes $n \in N$ and edges $e \in E$. Furthermore, nodes can have attributes $a_i \in A_{nodes}$, also, edges can have a number of attributes $a_j \in A_{edges}$. Each node $n \in N$ has associated attribute values v_i for each node attribute $a_i \in A_{nodes}$. Similarly, each edge $e \in E$ has associated attribute values v_j for each edge attribute $a_j \in A_{edges}$. Both the node and edge attributes can be ordinal (continuous or discrete) or categorical.

A predicate P_k over an attribute k has either the form $[v_{k_1}, v_{k_2}]$ (default $v_{k_1} = v_{k_{min}}$, $v_{k_2} = v_{k_{max}}$) if the attribute is ordinal (representing a range), or is of the form 0 (all) or v_k (single value) in case of a categorical attribute.

A selection of interest S_i now consists of a set of predicates $\{P_{k_i}\}$. To determine whether a node is in a selection, we use order of the selections to prevent conflicts, i.e., a node $u \in N$ always belongs to only one selection S_i:

$$u \in S_i \text{ if } u \notin S_j, j = 1 \ldots i-1 \text{ and } u.v_i \in \{P_{a_i}\} \forall a_i \in A_{nodes}. \qquad (1)$$

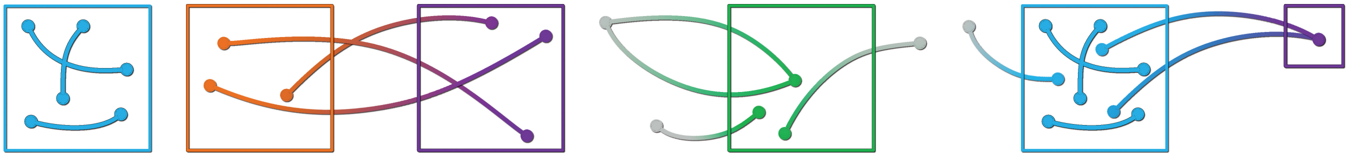
A node is contained in a selection if it is not already contained in a selection that is higher in the ordering, and its attribute values adhere to each of the selection predicates.

5.2 Interaction and Direct Manipulation

Following from the previously described selections of interest, we need to support three basic direct manipulation operations to enable visual querying:

- select current set,
- adapt range, and,
- change order.

For this we designed three different components in the graphical user interface; box visualizations representing the selections of interest



(a) Within. (b) Between. (c) Background. (d) Combined.

Fig. 4. Different types of edges involved in a node selection. a) Within edges showing all internal connections of a selection; both source and target node are contained in the selection. b) Between edges show all connections between two selections; both source and target node are contained in different selections. c) Background edges show all connections from a selection to the background selection. d) Combined, showing all involved edges for a selection, within, between and background.

in the detail view (see Figure 3a), an attribute component to adapt selection ranges (see Figure 3c), and a selection component to control the ordering (see Figure 3d).

In the attribute component, all node attributes are enlisted using Scented Widgets [45]. For each continuous attribute a scented span slider is shown and for each nominal or discrete ordinal attribute a scented selection widget is shown. In a different tab-page all edge attributes are shown, similar to the node attributes, using scented widgets. These attribute controls are directly linked to the current selection of interest. If a box is selected, either by point and clicking in the detail view or selection in the selection component, the scented widgets are updated to reflect the current attribute ranges or values. At any projection all attribute ranges can be adapted, also the ones currently not shown, to refine the current selection. Upon repositioning of the box in the detail view, the currently projected attribute ranges are updated.

Being able to slice-and-dice each attribute to refine the selection and directly gain feedback both on a structural- as well as the attribute-level provides users with a powerful exploration mechanism. For example, first two (or more) geographical regions of interest can be created in a `latitude-longitude` projection. Next, the projection is changed to `age` (x-axis) versus `income` (y-axis). Now the selection boxes can be freely repositioned and resized to slice-and-dice through the currently projected attributes while still maintaining the earlier defined geographic regions. Note that also dynamic network exploration is supported by being able to shift through time for both the nodes and edges if time is available as an attribute. See the video in the supplemental material for a demonstration of the different interaction methods.

5.3 Exploration

Next to the available attributes of nodes, additional derived attributes can be added based on a selection of interest. This enables, next to multivariate exploration, also exploration of the structure of the network. For example, in structural understanding it is interesting to find the nodes that are distance 1,2,3... etc. away from a certain node or group of nodes, or to identify the nodes that are not reachable from a certain group of nodes. For this, users can add a dynamic attribute that computes the distance (in terms of link hops) to a selection S_i. A value or range for this derived attribute can then be set to support structural exploration. The derived attributes are dynamically updated in real-time upon changing the associated multidimensional boundaries of the selection box by running Dijkstra shortest path algorithm [10] for all n nodes involved, having run time $\mathcal{O}(n \times |E|\log|E|+|V|)$.

For each selection consisting of nodes, there are three types of edges involved: *within*, *between* and *background* edges. For *within* edges both involved nodes are within the same selection, see Figure 4(a). Between edges have one node contained in one selection and the other in a different selection, see Figure 4(b). For background edges one node is contained in the selection and the other node is contained in the *background* selection, see Figure 4(c).

More formally, for a current selection of interest S_i an edge e with source and target nodes e_s and e_t respectively, has type e_{type}:

$$e_{type} = \begin{cases} \text{within} & \text{if } e_s \in S_i \text{ and } e_t \in S_i; \\ \text{between} & \text{if } e_s \in S_i \text{ and } e_t \notin S_i \text{ and } e_t \in S_j, j = 1 \ldots n; \\ \text{background} & \text{if } e_s \in S_i \text{ and } e_t \notin S_j, j = 1 \ldots n. \end{cases}$$

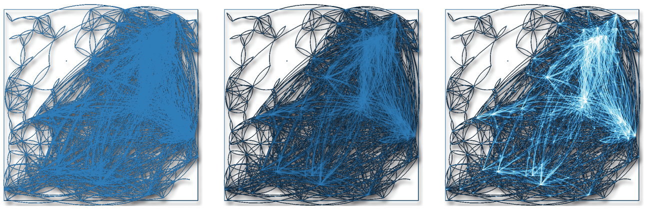
For between and background edges, we further distinguish between incoming and outgoing edges. Users can define for each selection which types of edges should be shown. Initially, for each new selection *within* and *between* edges are shown and *background* edges are hidden. Further filtering on the edges is supported similar to node filtering using Scented Widgets. Despite filtering options, there may be many edges involved for a selection, which clutters the view and prevents the identification of involved nodes and hides network structure. We improve upon this by introducing the option to enable transparent drawing of edges in combination with additive blending, see Figure 5.

Upon the creation of a new selection by adding a box to the current projection in the detail view, a linked aggregated visual representation is created automatically in the high-level infographic-style overview, discussed in the next section.

6 High-level infographic-style overview

The high-level infographic-style overview provides users with abstraction, insight, enables communication to a broader audience and is created semi-automatically based on the selections of interest. When a new selection box is created in the detail view, a linked visualization, sharing the same outline color, is added to the high-level overview.

The linked box shows aggregate information about the nodes in a selection. By default this is simply the number of nodes, shown textually, but also more detail can be shown. The visual representation can be changed to different multivariate visualization types. Currently we support scatter-plots, parallel-coordinate plots [22], histograms and (squarified) treemaps [8, 33]. For each of the visualizations the visual variables involved, such as what to show on the axes, can be changed to support further exploration and analysis (Figure 3f). We also support a small multiple exploration style similar to Van den Elzen et al. [38] in which multiple visualizations are created, one for each value of a visual variable, to enable comparison and guidance in the exploration.



(a) Standard rendering. (b) Opacity enabled. (c) Additive blending.

Fig. 5. A selection of interest containing many edges that clutter the view making it impossible to identify involved nodes and local structure. We improve upon (a) standard rendering by enabling (b) transparent drawing of lines and enabling (c) additive blending.

If the ranges of the selections of interest are updated then the associated visualizations in the overview are also updated automatically to reflect the changes. The visualizations can be freely positioned and resized using standard editing controls found in visual editing programs. We considered an automatic lay-out here, but since the number of selections is typically small (2-6) and because the structure is highly dependent on the semantics, we opted for supporting manual lay-out.

The visible edges in the detail view are aggregated and also shown in the high-level overview. *Within* edges are shown as self-loops of the associated visualization. *Between* edges are rendered between the associated selection visualizations and finally, *background* edges are drawn gradually semi-transparent to the background, not attached to anything. The width of the edges is proportional to the count or sum of a selected link attribute. Initially this is the number of edges. Users are enabled to show the actual values in textual form rendered on top of the edges. Aggregated edges in the overview can be filtered by setting a range using a scented span slider. For the color we use a gradient from the colors of start selection to the end selection.

Users are enabled to freely zoom and pan the high-level overview. Level-of-detail zooming is implemented for the visualizations; if users zoom in on a specific visualization or enlarge the visualization, more detail becomes available, for example, the name of the attributes shown on the axes. Visualizations are drawn semi-transparent to enable comparison of charts by (temporarily) overlaying one visualization on top of another.

7 Examples and Use cases

Below we describe some example DOSA explorations of real-world multivariate network data. We show how a tightly coupled exploration is achieved by starting with either multivariate data or the underlying network topology and show this in context of the other to find correlations, anomalies and patterns.

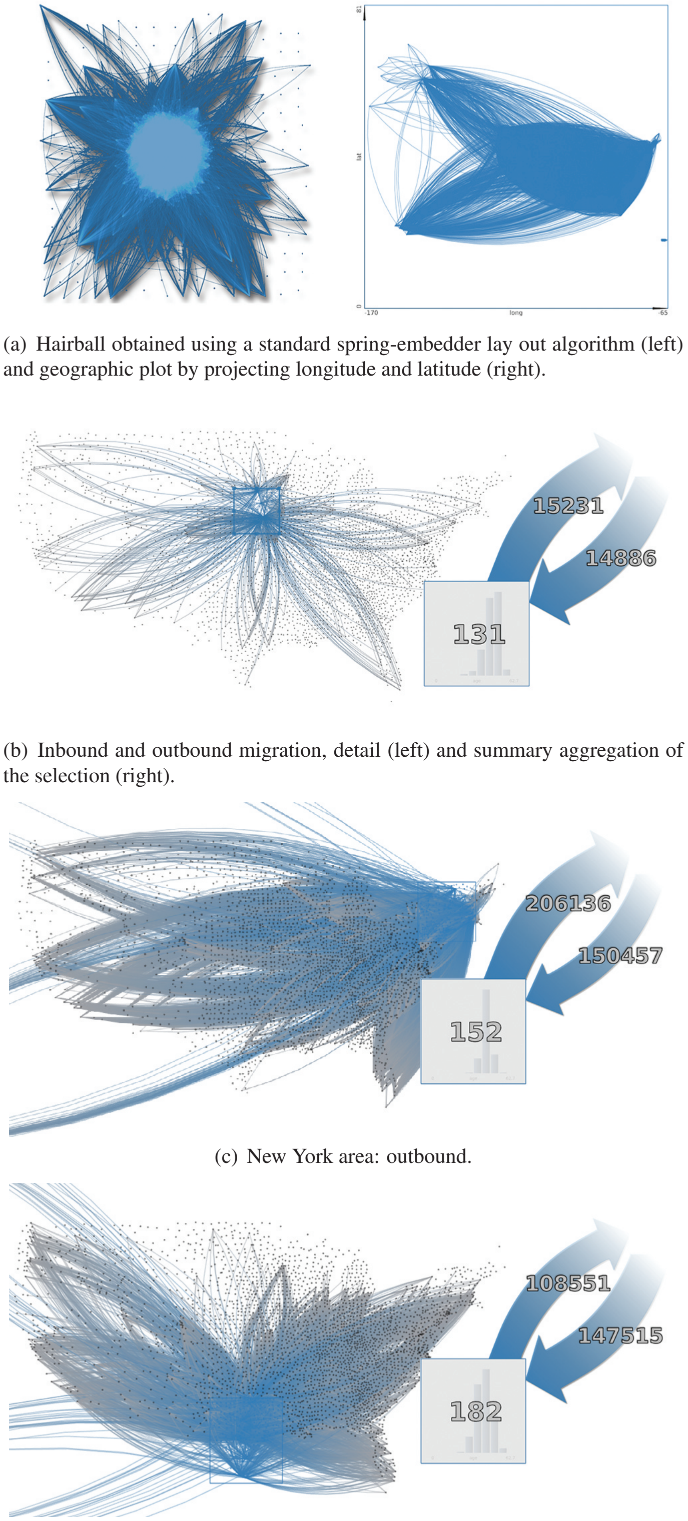
7.1 US Migration and Census

United States county to county migration data was obtained from year-to-year address changes reported on individual income tax returns filed with the IRS [1]. Next, this data was augmented with geographic location of the counties and according state and finally, combined with county census data provided by the United States Census Bureau [2]. The final dataset consists of 3,221 nodes (counties) and 78,294 edges (migrations), 14 node attributes and 10 edge attributes.

By using a standard spring-embedder algorithm [17] to lay out the nodes of the network, we hope to see structure, such as hubs, communities, and disconnected components. However, we are presented with a typical hairball-like visualization from which no insights can be gained and exploration is impossible, see Figure 6(a) left. Next, we switch to a more familiar geographical plot by using a `longitude`, `latitude` projection. We add a selection box to this projection that encloses all nodes, to see whether network structure is revealed, which unfortunately, is not the case, see Figure 6(a) right. Therefore, the box is resized to a smaller region to enable more focus.

7.2 Balance

We are interested in whether there are regions or states that are more inbound, outbound or balanced. To support this exploration, we disable *within* edges and enable *background* edges such that we can directly see the total incoming and outgoing aggregated migrations for our current selection box in the high-level overview, see Figure 6(b). Now, we can drag our box around in the detail view to quickly scan for unbalanced regions. We see that North-East, around the New York region, migration is more outbound, see Figure 6(c). In the South, the regions around Texas and Florida, migration is more inbound, see Figure 6(d), also Alaska is slightly inbound. If we resize the box and extend the selection to compare West with East, we find that both are balanced. If we next compare North with South, by adding another selection box, we find that North is more outbound and South is more inbound. By adding two more selection boxes we can refine the division of the United States into four regions. Now we see, Figure 7, that all migrations are balanced except for the North-East region; there, migration is more outbound to both South-East and South-West selections. We can conclude that North is more outbound due to people leaving from the North-West region. For the rest of the country migration is mostly balanced and no other anomalies are found.



(a) Hairball obtained using a standard spring-embedder lay out algorithm (left) and geographic plot by projecting longitude and latitude (right).

(b) Inbound and outbound migration, detail (left) and summary aggregation of the selection (right).

(c) New York area: outbound.

(d) Texas area: inbound.

Fig. 6. United States migration data exploration scanning for inbound, outbound and balanced regions.

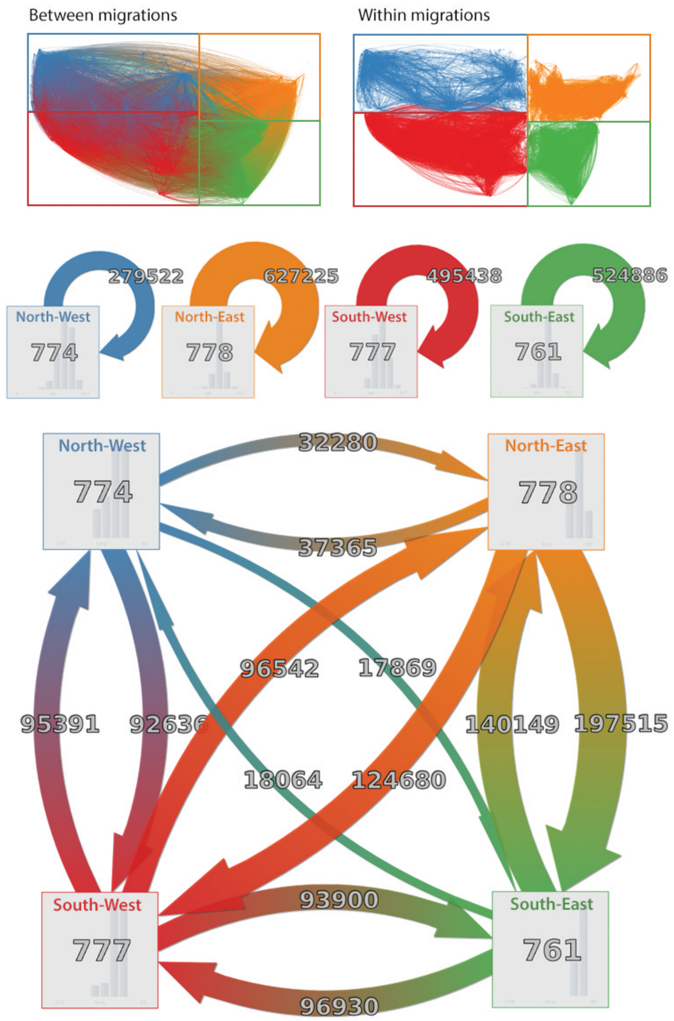


Fig. 7. United States migration data exploration testing for predominantly inbound or outbound regions with detail (top) and overview (bottom). Number of counties in each selection, shown on the boxes, is approximately equal to achieve fair comparisons. The North-East region is outbound with migrations mainly going to the South-East and South-West regions. The rest of the migrations are fairly balanced. The North-West has the least internal migrations and North-East the most (middle).

7.3 Crime

Next, we show how exploration of attribute and edge occurrence correlation is supported. For example, we want to investigate whether a high crime rate correlates with outbound migration:

- First, the nodes are positioned by selecting `crime-rate` and `population` as our current projection attributes, focusing on high population counties.
- We create two selections, dividing the top 25 highest population counties, one low crime-rate and the other high crime-rate.
- For both we enable *within* and *between* edges.

We now see that migration from high-crime counties to low-crime counties is higher than the other way around. We also see that migration within low crime areas is twice as high as within high crime areas. If we also enable outgoing background migration we see that significantly more people are leaving from high crime counties compared to low crime areas, see Figure 8.

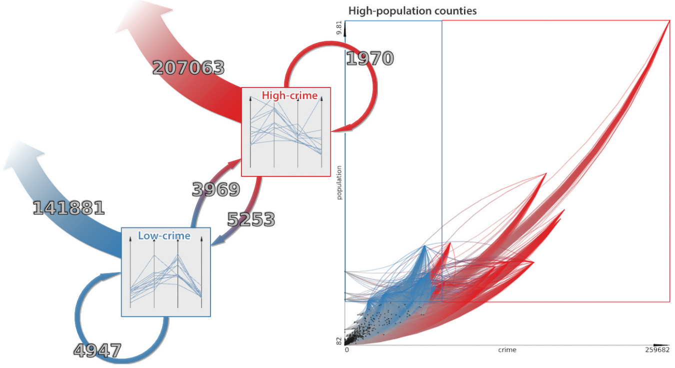


Fig. 8. Migration of highly populated low and high crime regions.

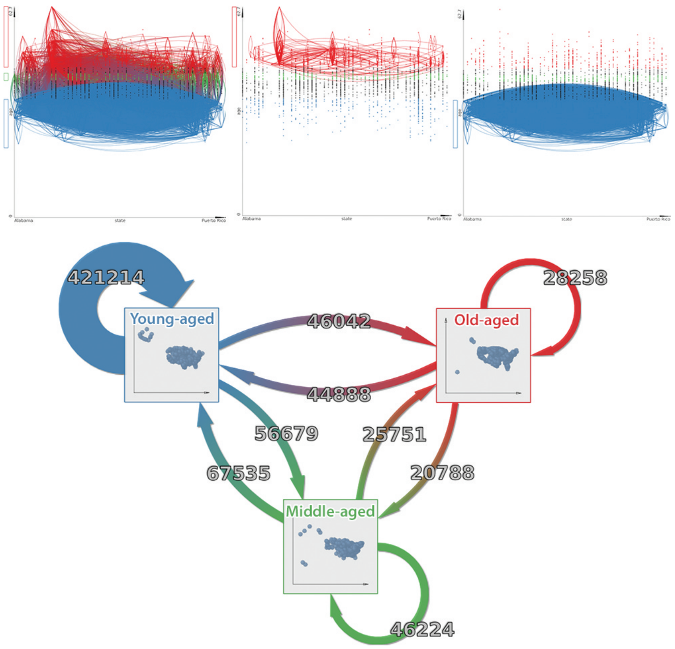


Fig. 9. Testing for correlation between age and migration.

7.4 Age

A similar investigation can be made for testing correlation between age and edge occurrence (migration). We project `state` versus `age` and create three selections: counties where average age is low, middle and high. We keep the number of counties roughly equal in each selection to be able to make fair comparisons. From the high-level overview we see that people from low-age counties tend to move to other low-age counties. We also see that they prefer to move to middle-age counties compared to high-age counties. People from low-age counties tend to move most, followed by people from middle-age counties and finally people from high-age counties. From the overview we also see in the geographical visualizations, that most part of Alaska contains young-aged persons, also some middle-aged but is not dominated by old-aged persons. Hawaii, however, does not contain dominantly young-aged counties. Finally, Florida mostly contains old-aged counties. Also, interesting observations from the detail view can be made, we already concluded that people from young-counties tend to migrate to other young-counties and people from old-counties tend to migrate to other old-counties. In the detail view we see that there is an additional pattern; people from young-counties move to other young-counties mostly in a *different* state, while people from old-counties mostly migrate to other old-counties *within* the same state. See Figure 9 for an overview.

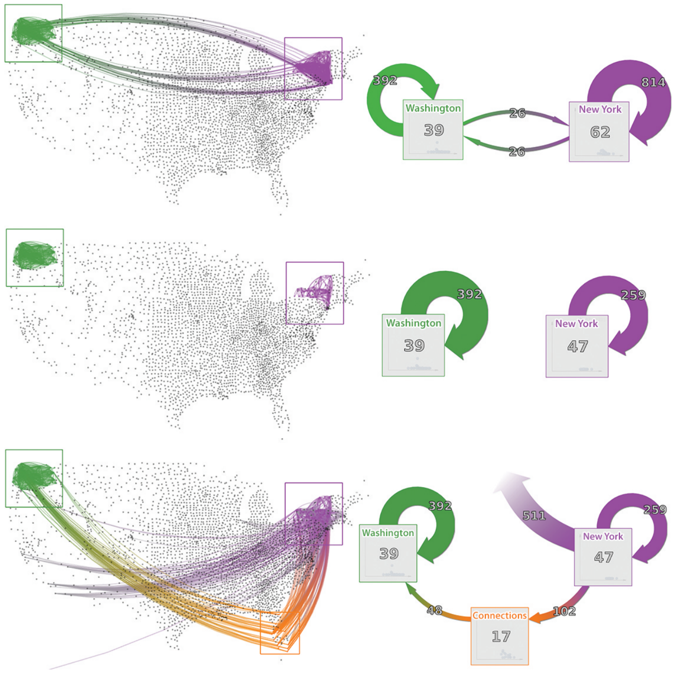


Fig. 10. Network path exploration, finding indirect routes from New York to Washington. Aggregated visual representations show number of nodes (counties) contained in them. Aggregated edges show number of edges. From New York there are 102 possible routes to 17 counties in the Florida region, from these 17 counties another 48 routes lead to Washington. From top to bottom we display: (top) Direct links (migrations) between New York and Washington, (middle) counties in New York that have exactly distance two to the counties selected for Washington, and (bottom) all outgoing links from the counties in the New York selection and a selection in the Florida region containing counties that connect New York with Washington.

7.5 Visual path discovery

Now, we show how structural exploration is supported with simultaneous multivariate data analysis. Assume, we are interested in finding patterns A to B, B to C. We want to find a region (B) where people from the state New York (A) are migrating to, and also people from this region are migrating to the state Washington (C). We introduce two selections in a geographical projection; one selection S_1 filtered on the state attribute New York, the other selection S_2 on Washington, both positioned to their according regions. Now we see the direct migrations between the two states, see Figure 10 (top). We introduce a new attribute, distance to New York, $D(S_1)$ and add this as a constraint to S_2, to only show nodes with distance 2 from S_1, see Figure 10 (middle). Next, we enable *background out* edges for S_2 and we are presented with all the counties via which people migrate from New York to Washington. We can now add another selection to be able to see the number of nodes and edges involved for one possible route, see Figure 10. This process can be repeated to find, for example, paths of length 3. Note that while performing these structural operations we can still filter on node and edge attributes per selection for combined multivariate analysis.

7.6 Enron Email corpus

All email traffic of Enron (former energy service company) corporation was made publicly available during the legal investigation of the biggest American bankruptcy due to accounting fraud [32]. The dataset is cleaned, private messages are removed, and, augmented with employee function. Furthermore, sentiment analysis was performed on the email body texts and added as multivariate link data. This dataset consists of 149 nodes (employees) and 185,506 edges (emails), 5 node attributes and 15 edge attributes.

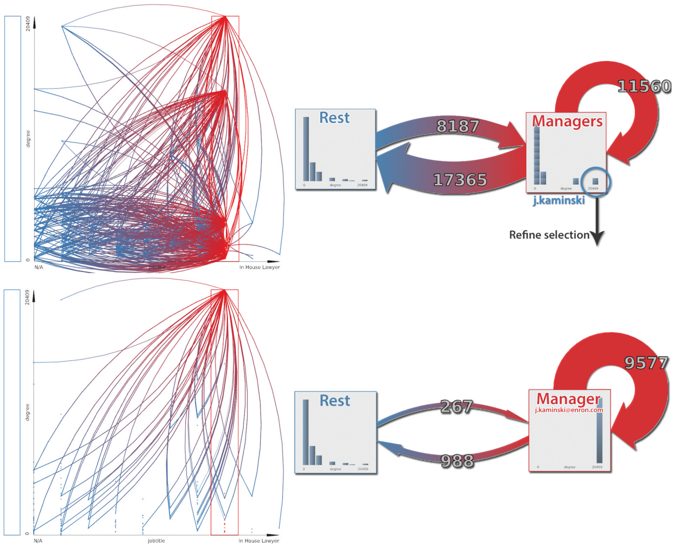


Fig. 11. Enron email communication exploration using two selections of interest: one representing the managers, the other the rest of the employees. Managers stand out due to a large self-loop (11,560 emails). After refinement the cause appears to be a single manager emailing himself all the time (9,577 emails).

Assume we are working at the human-resources department and want to explore the email behavior of our company. First we start by projecting the nodes according to `jobtitle` (x-axis) and `degree` (y-axis). Note that we are mixing multivariate data with a structural property here. This provides an overview of the distribution of employees in the different jobtitle groups and also who is emailing most within these groups. Next, we are going to explore the email behavior of the different groups. Therefore, we introduce two selections, one to select a specific group and the other containing the rest of the employees. For the latter selection we disable *within* edges to be able to focus on the *between* communication. We use the first group to shift through the different function groups. We see that:

- CEOs are more sending email,
- directors are more receiving,
- managers are heavily biased towards sending email, and,
- the managers stand out because they have a large self-loop in the overview.

From the overview visualization we see two persons having an unusual high degree. We identify the highest person via details on demand in the visualization and refine our selection to only contain this person, see Figure 11. Now it becomes clear, by refining the selection to show `cc` and `to` emails, that the high self-loop is indeed due to this person who cc-ed himself 9,422 times and directly sent himself another 155.

7.7 C-Level communication

Now assume we want to inspect CEO communication behavior. Therefore, we introduce another few selections and only show communication from and to the CEOs. We see that CEOs are mostly communicating with Vice presidents and managers. However, we also see that regular employees are communicating with CEOs, which is not what we would expect. By refining our selection we see that it is only one person, Jeff Dasovich, who is heavily communicating with the CEOs and mainly broadcasting, see Figure 12 (left). By filtering on the sentiment attributes we also see that his emails are mainly negative. After googling it appears that Jeff Dasovich is Enron's Government Relation Executive, who had to communicate to the CEOs when

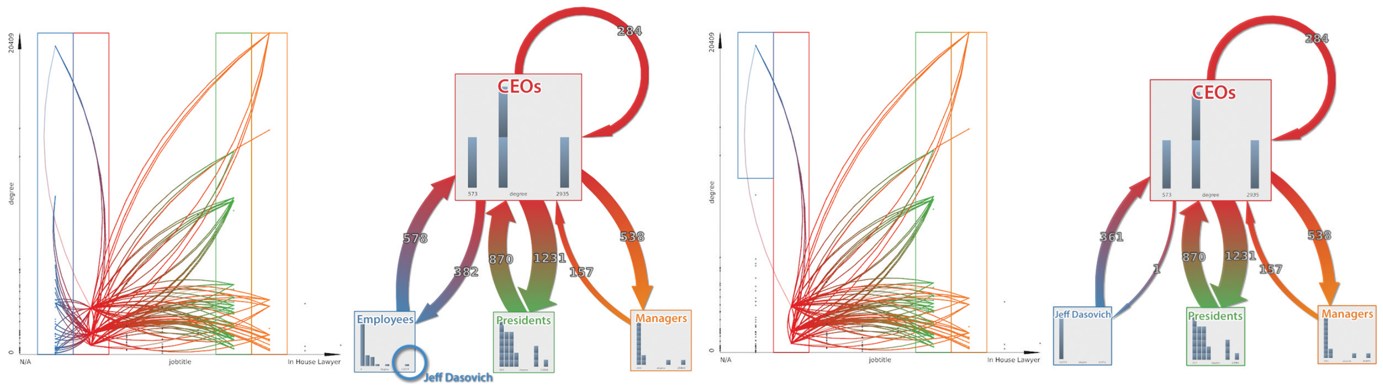


Fig. 12. Typical C-level communication: CEOs are heavily communicating with Vice presidents and managers. However, also communication is present between CEOs and regular employees, which turns out is only one person heavily broadcasting to the CEOs (right).

things went wrong for Enron, mislabeled here as a regular employee. If we now keep this configuration and refine the selection to contain only emails with strong problem sentiment in it, we see it contains only one-way communication of the vice-presidents and Jeff Dasovich to the CEOs, see Figure 12 (right). We refine the selection again on time and see that these emails were mostly sent during the critical period for Enron.

8 Discussion and Limitations

The basic ideas presented in this paper are all simple in nature: 1) combined structural and multivariate exploration and analysis through visual queries using 2) selections of interest, based on (derived) node and link attributes, controlled by Scented Widgets and playful interaction on two 3) juxtaposed and linked views showing network detail using (derived) attribute projections and showing a high-level infographic-style summary overview simultaneously. However, combined they are novel and enable a strong visual query mechanism that is intuitive and effective.

In an earlier stage of the research, the prototype only contained a single view: the detail view. The aggregated visualizations in the overview were directly rendered on top of the selections of interest. However, this had several disadvantages; upon dragging the selection it was difficult to track changes in the visualization due to the motion and also internal edges of a selection where no longer visible due to the overlapping visualization. Therefore, it was decided that two juxtaposed views would benefit users in the exploration and analysis. This enabled also the possibility to have the easy to understand high-level overview as a means of presentation and communication to people who are not interested in the low-level details.

We also consider it a strong point of our system that it is simple in design and not much explanation is needed. Quite some effort is put into keeping interaction intuitive, uniform and minimalistic, e.g., similar interaction methods in both views (select, drag, change size), similar interaction to deal with node and edge attributes (uniform Scented Widgets), uniform and combined interaction for structural and multivariate exploration based on attributes. Also, visual coherence between the different components is achieved by using color. Visual elements are kept to a minimum to reduce visual noise and support a broad audience.

There are, however, some limitations such as scalability with respect to number of attributes and scalability with respect to number of selections. We believe that in general users are content with about 5 or at most 10 selections of interest in order to answer their questions and still being able to understand the involved complexity. But if the number of selections of interest becomes large, for example due to automatic clustering or community detection algorithms, then both the detail and overview become cluttered. In the overview, edges below a certain threshold can be filtered as well as in the detail view, however, this is only partly a solution.

Currently we are relying on a node partitioning and nested or overlapping selections of interest are not possible. Enabling this allows for more powerful queries but also increases interpretation difficulty and selection mechanism conflicts. Also, since we are relying on ranges as multidimensional boundaries for the selections, we only allow for box-shaped selection widgets. Brushing capabilities could be introduced such that users can freely color nodes for a certain selection. However, the uniform selection mechanism based on scented widgets would break, as now nodes are no longer identifiable by range but only based on unique identifier. This implies that potentially a large number of gaps appear in the attribute ranges making interaction of the Scented Widgets a challenge.

If the number of attributes associated with the nodes and edges is large, the list of Scented Widgets becomes difficult to interact with due to a large scrolling area. Also, the number of available projections for the detail view quickly grows (n^2, for n attributes) making it more likely that interesting features in the data are missed. A solution here could be feature selection, to only select the most interesting features, as a preprocessing step of the data before loading it into the tool either automatic or manually using visualization techniques [5, 13, 29].

9 Conclusions

We presented novel interaction methods for both domain-expert and casual users to explore and analyse multivariate networks concurrently on network topology as well as the multivariate data. This enables users to see outliers, patterns and trends for the combined elements. Furthermore, we support users in the simultaneous creation of a high-level infographic-style overview. This helps in understanding the network due to a simple image, provides abstraction and aggregation and presents a means for communication to a broader audience. Both interaction methods are facilitated by using selection sets as a central element and the juxtaposition of detail and overview. We have shown the effectiveness of our DOSA approach through several elaborated examples on real-world datasets. Furthermore, we have shown this method is not just limited to multivariate networks, but also functions when only multivariate data or network structure is available. Finally, due to the general and flexible setup, this method is domain-independent.

9.1 Future work

For future work it is interesting to enable the partitioning of the network not only on the nodes but also on the edges. This could enable richer exploration by showing aggregate visualizations also for the edges. However, intuitive interaction and facilitation of this is not trivial. Also, the adaption of the interaction methods would benefit from the enhancement of the detail view to support different visualizations such as matrix visualization or ultimately generalize the techniques to any network visualization. Finally, functionality could be added to export the high-level infographic-style overview to an external editing tool such as Microsoft PowerPoint or Adobe Illustrator for further fine-tuning, editing and enrichment, e.g., for publication.

References

[1] I.R.S. Internal Revenue Service, SOI Tax Stats - County-to-County Migration Data Files. http://www.irs.gov/uac/SOI-Tax-Stats-County-to-County-Migration-Data-Files. Accessed March, 2014.

[2] United States Census Bureau, 2010 Census Data. https://www.census.gov/2010census/data/. Accessed March, 2014.

[3] D. Archambault, T. Munzner, and D. Auber. GrouseFlocks: Steerable Exploration of Graph Hierarchy Space. *IEEE Trans. Vis. Comput. Graphics*, 14(4):900–913, 2008.

[4] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. An Alan R. Apt Book. Prentice Hall, 1999.

[5] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.-D. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. In *Proc. 12th Eurographics IEEE Conf. Visualization*, pages 863–872. Eurographics Association, 2010.

[6] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and Magic Lenses: The See-through Interface. In *Proc. 20th Annu. Conf. Computer Graphics and Interactive Techniques*, pages 73–80. ACM, 1993.

[7] L. Borisjuk, M. reza Hajirezaei, C. Klukas, H. Rolletschek, and F. Schreiber. Integrating Data from Biological Experiments into Metabolic Networks with the DBE Information System. In *Silico Biology 5: 0011*, page 11, 2005.

[8] M. Bruls, K. Huizing, and J. J. van Wijk. Squarified Treemaps. In *Proc. Joint Eurographics and IEEE TCVG Symp. Visualization*, pages 33–42. Press, 1999.

[9] G. Colliat. OLAP, Relational, and Multidimensional Database Systems. *SIGMOD Rec.*, 25(3):64–69, 1996.

[10] E. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[11] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson. GraphTrail: Analyzing Large Multivariate, Heterogeneous Networks While Supporting Exploration History. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 1663–1672. ACM, 2012.

[12] C. Dunne and B. Shneiderman. Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 3247–3256. ACM, 2013.

[13] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1141–1148, 2008.

[14] N. Elmqvist, J. Stasko, and P. Tsigas. DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data. In *Proc. IEEE Symp. Visual Analytics Science and Technology*, pages 187–194, 2007.

[15] J.-D. Fekete, D. Wang, N. Dang, and C. Plaisant. Overlaying Graph Links on Treemaps . IEEE Symp. Information Visualization Conf. Compendium (demonstration), 2003.

[16] F. Frasincar, A. Telea, and G.-J. Houben. Adapting Graph Visualization Techniques for the Visualization of RDF Data. In V. Geroimenko and C. Chen, editors, *Visualizing the Semantic Web*, pages 154–171. Springer London, 2006.

[17] E. R. Gansner and S. C. North. An Open Graph Visualization System and its Applications to Software Engineering. *Software- Practice and Experience*.

[18] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist. Visual Analytics for Multimodal Social Network Analysis: A Design Study with Social Scientists. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2032–2041, 2013.

[19] I. Herman, G. Melancon, and M. Marshall. Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE Trans. Vis. Comput. Graphics*, 6(1):24–43, 2000.

[20] D. Holten. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Trans. Vis. Comput. Graphics*, 12(5):741–748, 2006.

[21] C. Hurter, B. Tissoires, and S. Conversy. FromDaDy: Spreading Aircraft Trajectories Across Views to Support Iterative Queries. *IEEE Trans. Vis. Comput. Graphics*, 15(6):1017–1024, 2009.

[22] A. Inselberg. The Plane with Parallel Coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[23] I. Jusufi, Y. Dingjie, and A. Kerren. The Network Lens: Interactive Exploration of Multivariate Networks Using Visual Filtering. In *Proc. 14th Int. Conf. Information Visualisation (IV)*, pages 35–42, 2010.

[24] I. Jusufi, A. Kerren, and B. Zimmer. Multivariate Network Exploration with JauntyNets. In *Proc. 17th Int. Conf. Information Visualisation (IV)*, pages 19–27, 2013.

[25] A. Kerren, H. C. Purchase, and M. O. Ward, editors. *Multivariate Network Visualization*. Number 8380 in Lecture Notes in Computer Science. Springer International Publishing, 2014.

[26] A. Martin and M. O. Ward. High Dimensional Brushing for Interactive Exploration of Multivariate Data. In *Proc. IEEE Conf. Visualization*, page 271, 1995.

[27] A. Perer and B. Shneiderman. Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Trans. Vis. Comput. Graphics*, 12(5):693–700, 2006.

[28] A. J. Pretorius and J. J. van Wijk. Visual Inspection of Multivariate Graphs. *Comput. Graph. Forum*, 27(3):967–974, 2008.

[29] J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data. *Information Visualization*, 4(2):96–113, 2005.

[30] A. Shamir and A. Stolpnik. Interactive Visual Queries for Multivariate Graphs Exploration. *Computers & Graphics*, 36(4):257–264, 2012. Applications of Geometry Processing.

[31] Z. Shen, K.-L. Ma, and T. Eliassi-Rad. Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. *IEEE Trans. Vis. Comput. Graphics*, 12(6):1427–1439, 2006.

[32] J. Shetty and J. Adibi. The Enron Email Dataset Database Schema and Brief Statistical Report. http://www.cs.cmu.edu/∼enron/, 2004.

[33] B. Shneiderman. Tree Visualization with Tree-maps: 2-d Space-filling Approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.

[34] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. IEEE Symp. Visual Languages*, pages 336–343, 1996.

[35] B. Shneiderman and A. Aris. Network Visualization by Semantic Substrates. *IEEE Trans. Vis. Comput. Graphics*, 12(5):733–740, 2006.

[36] C. Tominski, J. Abello, and H. Schumann. CGV - An Interactive Graph Visualization System. *Computers & Graphics*, 33(6):660–678, 2009.

[37] C. Tominski, J. Abello, F. van Ham, and H. Schumann. Fisheye Tree Views and Lenses for Graph Visualization. In *Proc. Int. Conf. Information Visualisation*, pages 17–24, 2006.

[38] S. van den Elzen and J. J. van Wijk. Small Multiples, Large Singles: A New Approach for Visual Data Exploration. *Comput. Graph. Forum*, 32(3pt2):191–200, 2013.

[39] F. van Ham and A. Perer. Search, Show Context, Expand on Demand: Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Trans. Vis. Comput. Graphics*, 15(6):953–960, 2009.

[40] J. J. van Wijk and W. A. Nuij. A Model for Smooth Viewing and Navigation of Large 2D Information Spaces. *IEEE Trans. Vis. Comput. Graphics*, 10(4):447–458, 2004.

[41] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. Fellner. Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. *Comput. Graph. Forum*, 30(6):1719–1749, 2011.

[42] M. O. Ward. XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. In *Proc. IEEE Conf. Visualization*, pages 326–333, 1994.

[43] M. O. Ward. A Taxonomy of Glyph Placement Strategies for Multidimensional Data Visualization. *Information Visualization*, 1(3-4):194–210, 2002.

[44] M. Wattenberg. Visual Exploration of Multivariate Graphs. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, CHI '06, pages 811–819. ACM, 2006.

[45] W. Willett, J. Heer, and M. Agrawala. Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE Trans. Vis. Comput. Graphics*, 13(6):1129–1136, 2007.

[46] N. Wong, M. S. T. Carpendale, and S. Greenberg. EdgeLens: An Interactive Method for Managing Edge Congestion in Graphs. In *Proc. 9th Annu. IEEE Conf. Information Visualization*, pages 51–58, 2003.

[47] Y. Wu and M. Takatsuka. Visualizing Multivariate Network on the Surface of a Sphere. In *Proc. Asia-Pacific Symp. Information Visualisation*, pages 77–83. Australian Computer Society, Inc., 2006.