GraphTrail: Analyzing Large Multivariate, Heterogeneous Networks while Supporting Exploration History

Cody Dunne[1,2], Nathalie Henry Riche[1], Bongshin Lee[1], Ronald A. Metoyer[3], George G. Robertson[1]

[1]Microsoft Research
One Microsoft Way
Redmond, WA, USA
{nath, bongshin}
@microsoft.com

[2]Human-Computer Interaction Lab
Department of Computer Science
University of Maryland
College Park, MD, USA
cdunne@cs.umd.edu

[3]School of Electrical Engineering
and Computer Science
Oregon State University
Corvallis, OR, USA
metoyer@eecs.oregonstate.edu

ABSTRACT

Exploring large network datasets, such as scientific collaboration networks, is challenging because they often contain a large number of nodes and edges in several types and with multiple attributes. Analyses of such networks are often long and complex, and may require several sessions by multiple users. Therefore, it is often difficult for users to recall their own exploration history or share it with others. We introduce GraphTrail, an interactive visualization for analyzing networks through exploration of node and edge aggregates that captures users' interactions and integrates this history directly in the exploration workspace. To facilitate large network analysis, GraphTrail integrates aggregation with familiar charts, drag-and-drop interaction on a canvas, and a novel pivoting mechanism for transitioning between aggregates. Through a three-month field study with a team of archeologists and a qualitative lab study with ten users, we demonstrate the effectiveness of our design and the benefits of integrated exploration history, including analysis comprehension, insight discovery, and exploration recall.

Author Keywords

Network visualization; visual analytics; data aggregation; exploration history; analytic provenance.

ACM Classification Keywords

H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6).

General Terms

Design

INTRODUCTION

Networks like social networks and scientific collaborations offer rich sources of information. They are often large, ranging from several thousands to millions of nodes and edges. In addition, these network items often have multiple attributes (i.e., multivariate) and can include of several types of nodes and edges (i.e., heterogeneous). The large size and complex structure of these networks make it challenging to explore them, and developing tools to do so has been an active research topic (e.g., [2,16,32]). Traditional node-link diagrams have advantages, but do not scale well and often produce cluttered overviews. Various approaches for dealing with these networks exist, such as multiple coordinated views of ordered lists and histograms [16,18]. PivotGraph [32], alternatively, aggregates nodes by their attributes and places the node aggregates on an attribute-based grid. Aggregate edges between the node groups are shown with arcs.

We advance these aggregation approaches in GraphTrail (Figure 1), an interactive visualization system that supports exploration of large multivariate and heterogeneous networks. GraphTrail allows users to aggregate according to both node and edge attributes and to iteratively explore the aggregates. To aid understanding, GraphTrail employs familiar charts such as bar charts or tag clouds and lets users seamlessly switch to charts that best support their tasks. Users arrange charts on an interactive canvas and explore using drag-and-drop interactions. To filter, users select aggregates, drag them to the canvas to create a new chart, and choose a chart type to represent it. They can drag aggregates from one or more charts into an existing one to compose meaningful subsets. GraphTrail also offers a novel pivot mechanism to move from a subset to any connected set along a linking edge type. Analyses with GraphTail may require multiple sessions and result in several exploration branches, which users can merge to combine results.

Following the steps of a complex analysis can be difficult, but providing analysts with rich exploration histories can enhance their recall between sessions [31,21,29]. Current visualization systems usually show exploration histories in additional panels, and many require extra effort from users to maintain. Also, few history models support the complex explorations GraphTrail allows. In GraphTrail, each chart represents an aggregated subset of the network and possibly illustrates a finding. The core interaction we provide is the drag-and-drop refinement of aggregates, which can be captured and displayed directly in the workspace without additional user effort or history views. Exposing the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2012, May 5-10, 2012, Austin, TX, USA.
Copyright 2012 ACM 978-1-4503-1015-4/12/05...$10.00.

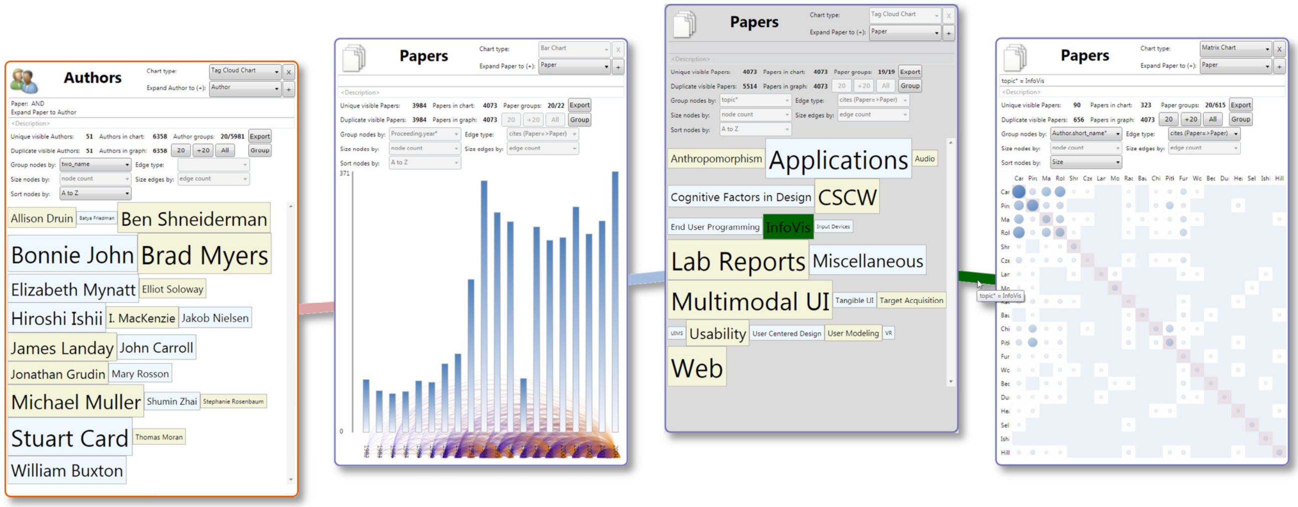

Figure 1. Four overviews of the CHI dataset showing (left to right): prolific authors, papers by year, popular topics, and InfoVis citation patterns.

analysis process in this way enables users to utilize their spatial memory while visual and textual feedback helps them track their interactions and share results with others.

We report the results of a three-month deployment of GraphTrail to a team of archeologists analyzing their own data. We then describe a qualitative lab study assessing how novices use the system for exploration. The results of these studies demonstrate that users can successfully use GraphTrail to analyze large multivariate and heterogeneous networks, and that our design helps exploration recall and analyses sharing. Our studies showed that with GraphTrail users were able to understand the analyses of others without annotations, which may be a first step towards enabling asynchronous exploratory analysis. Specifically, the contributions of this paper are:

- A system for exploring large multivariate, heterogeneous networks using aggregation by node and edge attributes,
- A method for capturing a user's exploration history and integrating it directly into the workspace, and
- A longitudinal field study and a qualitative lab study that prove the utility of these approaches.

RELATED WORK

Network analysis systems generally focus on node-link diagrams, as in Guess [3], Pajek [4] and SocialAction [23], or matrix representations, like in Matrix Zoom [1] and MatrixExplorer [12]. Both show the topology of small networks well but can be unreadable with a few thousand nodes. One way to get an overview of a large network is to aggregate nodes based on the topology. For example, Ask-GraphView [2] aggregates densely linked nodes together into meta-nodes. While this tactic can show the aggregated topology of networks with hundreds of thousands of nodes, it does not offer much support for exploration of attributes.

Multivariate networks, such as social networks, are particularly interesting and challenging to analyze. In node-link and matrix representations, multiple attributes can be encoded using size, color, shape, or opacity [7]. However, it is still challenging to identify patterns and extract trends with only these visual encodings. Various hybrid network visualizations attempt to combine topology and multivariate data into a single visualization (e.g., nodes in a scatterplot connected by edges [5,28]). Other hybrid approaches provide topology visualization on top of node aggregates (overlaying edges on Treemaps [9], combining Treemaps with node-link diagrams [33], or using matrix charts within aggregate node-link charts [13]). It remains challenging to analyze large multivariate networks with these visualizations because they lack support for exploration history and multi-session analysis.

Several recent attempts have produced promising results. For example, ManyNets [10] partitions networks according to attributes or topological properties, supporting easy comparisons of partition statistics. However, it is difficult to extract patterns and to identify relationships between attributes. In contrast, PivotGraph [32] aggregates nodes by attributes and shows relationships between aggregates using arcs, but does not support comparisons of more than two attributes or multiple types of network items.

Multivariate networks that are also heterogeneous pose an additional challenge for visual analytics that has been tackled by tools such as FacetLens [19] and NetLens [16]. FacetLens can view networks with several types of nodes and many attributes, grouping nodes by attribute values (facets) and pivoting between node types. However, users can only pivot from a single node to linked nodes. FacetLens helps users see patterns in node attributes, but does not show network topology. NetLens is well suited for content-actor networks with two node types. It uses two coordinated views where each view contains nodes aggregated by their attributes. Users explore networks by filtering in one view and pivoting from their filtered subset to connected nodes in the other view. NetLens allows for complex analysis scenarios and extraction of patterns in multivariate content-actor networks, but is limited to only two node types. In addition, complex filtering and pivoting operations may be difficult for users to track, especially when resuming a previous analysis.

Though the final visualizations created by an analyst demonstrate the findings they discovered, understanding the analytical process that produced those findings has been recognized as equally important by the visual analytics community [22]. Kang and Stasko [17] recommend that visual analytics systems support asynchronous collaboration for exploratory analysis, and providing easily understandable exploration histories can aid in this effort. Other studies have shown the benefit of recording and showing exploration histories. In a wire transfer analysis tool, exploration history views served as an effective mental aid for user strategies and choices, and boosted result confidence [21]. Users of a scatterplot visualization tool with exploration histories reported that recording their findings, linking them to visualizations, and organizing them were very important to the exploration process [29]. Also, Ware *et al.* [31] report a user study of a network analysis tool where task recall (after a week) improved 22% to 66% when using history traces, and the traces also encouraged more efficient search strategies.

Most existing visualization tools show exploration history as a list of actions that users can revisit (e.g., NetLens [16]) or a linearized history using thumbnails of the states (e.g., [11,14]). Linear histories do not capture all branches and intermediate states of an exploration, making it difficult for analysts to retrace the steps leading to a finding. Alternative approaches visualize exploration history as a tree [29,15]. Few history models support the non-linear and non-hierarchical exploration paths that analysts typically follow. Notably, Exbase [20] supports merging subsets but the exploration history is not exposed to users and, like [15], it requires users to replay scripts to reach a specific point in the history. The P-Set history model [14] supports merging branches, but none of its implementations directly present the history to users. More effective visual analytics systems would combine these techniques, allowing the manipulation (e.g., merge) of branches in the history and displaying the exploration history directly to users.

Moreover, most history models do not capture both interactions and transient states. Because states are often not displayed or displayed as static thumbnails, users are unable to perceive the chain of actions that led to visualizations, much less interact with or combine the results of analysis branches. Also, these history features are not integrated with the exploration process and require additional panels such as the workflows in VisTrails [5]. Annotations added to visualizations or a thumbnail history (e.g., [29,11,30]) help analysts describe their motivation or process, but require substantial user effort and diligence.

In summary, despite extensive efforts network visualization remains a challenge for large multivariate, heterogeneous networks. Existing systems either provide a simplified two-node-type model or treat node types equally. We designed GraphTrail to analyze these networks via node and edge aggregation and iterative refinement of the aggregates. Additionally, our design exposes transient states as separate charts on a canvas, integrating the user's exploration history directly within the workspace. We enhance the value of providing history views to users with visual and textual feedback to help users track interactions and share results.

GRAPHTRAIL

To aid understanding, GraphTrail represents aggregated views of the network using familiar charts instead of node-link diagrams. It includes tag clouds, matrices, tables, and a hybrid bar chart (Figure 2) we designed to gain an overview of aggregate node counts and the aggregate edge counts connecting them. This hybrid bar chart is similar to that of Pretorius and van Wijk [24]. The system can be easily

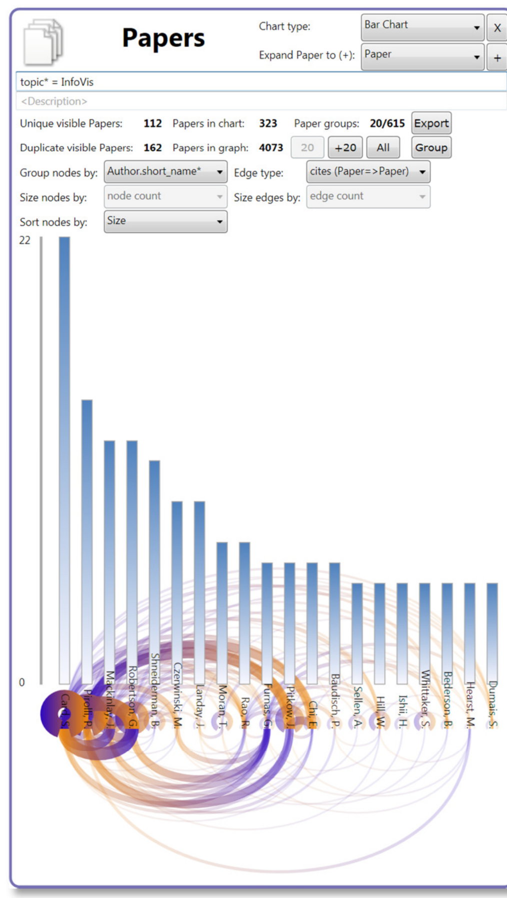

Figure 2. Hybrid bar chart showing CHI papers with the InfoVis topic. Bars show papers aggregated by their authors and the arcs show the number of citations between paper aggregates.

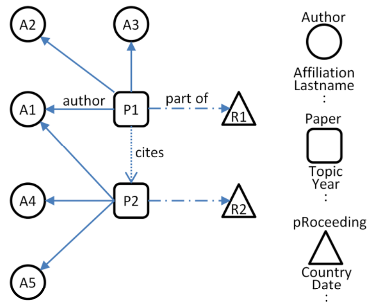

Figure 3. The CHI network model has author, paper, and proceeding nodes. Edges show citations between papers and connect papers to their authors and proceedings.

extended to include other types of charts as well. GraphTrail is designed to show exploration history by recording a subset of meaningful actions the user takes and displaying them as visual and textual cues on an infinite canvas. The primary interactions involve dragging and dropping subsets of data onto the canvas and choosing the desired chart type. Users organically arrange the analysis as they create it and focus on specific parts using zooming and panning. The history is exposed at all times, and users can interact with or combine any previous charts they created.

We illustrate GraphTrail's interface with a scenario inspired by NetLens [16]. In our discussion, we make reference to components that will be explained in detail in the design section. The scenario is based on exploring the CHI papers from 1982 to 2004, which consists of 4,073 papers linked to their authors and proceedings. Papers link to each other via citations and authors to each other via co-authorship. Example paper attributes are title and a manually annotated topic, and author attributes include name and affiliation. Figure 3 shows a subset of the network model. This dataset has three types of nodes: papers, authors, and proceedings. However, GraphTrail is not limited to three node types.

Scenario: Learning about Groups of Researchers

Let us imagine that Emma, a student looking for research opportunities, uses GraphTrail to investigate HCI research in two states: Georgia and Washington. Emma loads the CHI paper dataset and GraphTrail presents her with a hybrid bar chart (Figure 4, ROOT), showing papers grouped by topic and sorted by size, as well as the aggregated citation pattern linking topics. She finds it interesting that InfoVis papers tend to cite each other, as shown by the thick self-loop circle for the topic (triangle).

However, Emma is most keen on learning about individual researchers, so she first pivots to the authors of the papers using the interface (star). GraphTrail creates a new bar chart showing the authors connected to these papers by the authorship edge type. Since she is interested in authors from Georgia, she groups the nodes by "state," updating the chart in place (Figure 4, Chart 1). From this chart she drags the

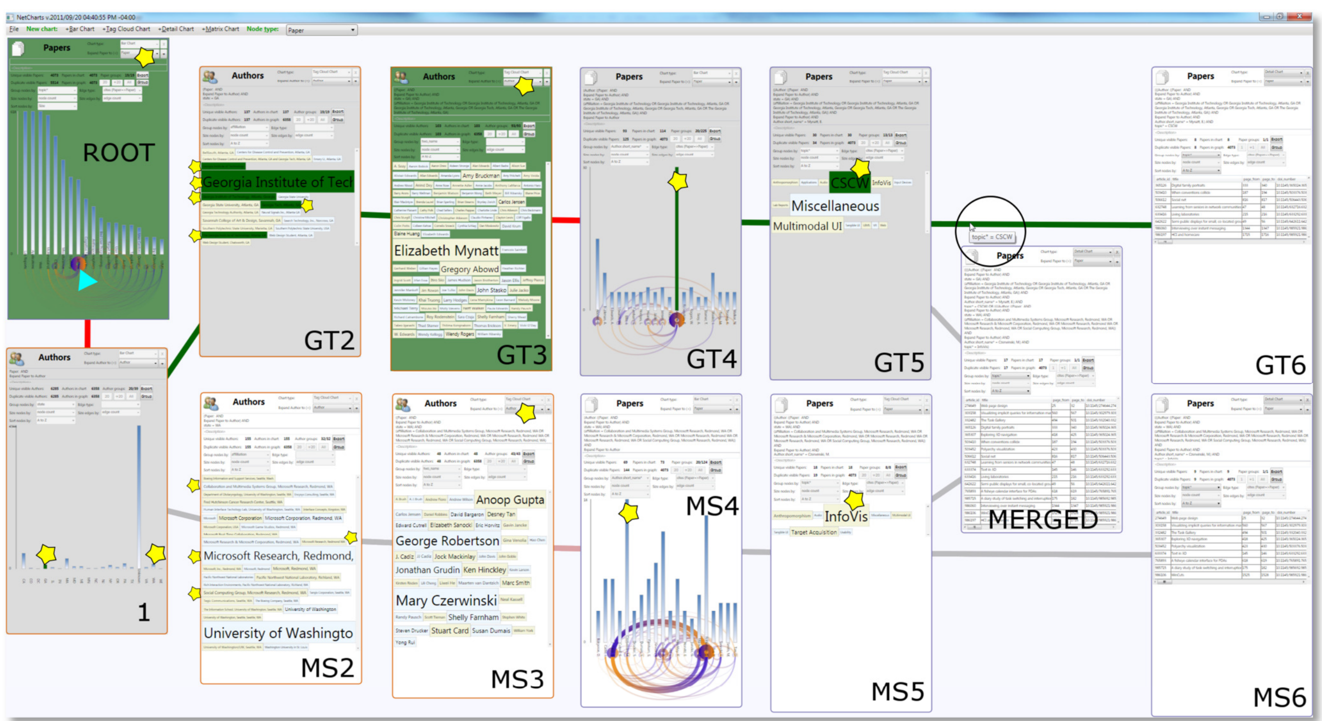

Figure 4. A GraphTrail analysis showing two parallel exploration paths, the top examining Georgia Tech (GT) patterns and the bottom comparing Microsoft Research (MS). They start at the ROOT chart that contains all the papers in the dataset. Charts in each path are numbered in order of creation (e.g., 1, GT2, GT3, …), and the user interactions are shown with stars. The MERGED chart is the union of both branches' results. The user moved the mouse over the final parent link in the GT path (circled), highlighting the chain of actions up to the root.

bar for “GA” (left star) out of the chart and drops it onto the canvas to filter. GraphTrail creates a chart with only the selected subset. To focus on the impact of the institutions in Georgia, Emma aggregates the authors by affiliation and changes the size encoding to reflect the number of papers from each. As the vertical labels of a bar chart are hard to read, she changes the chart to a tag cloud (Figure 4, GT2), with the affiliations sorted alphabetically.

In the resulting chart (Figure 4, GT2), Emma identifies the most prolific institution as Georgia Tech, but she also notices that there are five name variations of Georgia Tech in the tag cloud (stars). To get an accurate understanding, Emma merges these five subsets together. She drags the first Georgia Tech tag and drops it onto the canvas to create a new tag cloud chart. She then drags each of the four remaining tag variations out and drops them onto this new chart. To visualize the names of the researchers of this institution, Emma groups the nodes by “name” (GT3).

This chart (Figure 4, GT3) shows the most prolific authors affiliated with Georgia Tech, the font size encoding their number of papers. She notices that Elizabeth Mynatt has the most. Emma is now interested in exploring Mynatt’s work in more detail in the context of the Georgia Tech papers. She pivots to the papers written by these Georgia Tech authors, selects the hybrid bar chart type, and groups the papers by author (GT4). To see the topics of interest to Mynatt, Emma drags the bar for Mynatt (star) out and drops it onto the canvas, choosing a tag cloud and grouping by topic (GT5). Interested in the CSCW topic, Emma uses the drag and drop filtering to see Mynatt’s papers published in this area in a table view (GT6). Intrigued by three of these papers, Emma downloads them for her afternoon reading.

Emma repeats a similar process to compare the state of Washington, creating a parallel path below the first (Figure 4, MS2-6). She focuses on Microsoft Research, specifically Mary Czerwinski’s work in the InfoVis topic. Finally, she merges the results of the GT and MS branches (MERGED).

When she returns three days later to add Massachusetts to her analysis, she quickly follows each path left to right. When she moves the mouse over a link between two charts, GraphTrail provides visual and textual feedback showing her history of actions. In Figure 4, her mouse (circled) is over the parent link of chart GT6, which is highlighted in green and shows a tooltip for the filtering action. Also, each aggregate she dragged into a subsequent chart is highlighted (stars in GT2, GT4, and GT5). For each pivot, the entire chart is highlighted and parent links are shown in red. This helps her quickly recall how she created each chart.

GraphTrail Design

Data Aggregation

GraphTrail’s design strategy is to handle large networks by aggregating nodes and edges instead of showing them individually in a node-link diagram. Consider the example network model shown in Figure 3. GraphTrail aggregates network items by attributes to produce an aggregate chart (e.g., a bar chart). In Figure 3, grouping papers by topic would yield a bar chart of all papers organized by topic.

GraphTrail allows nodes to have multiple values for a single attribute (e.g., papers have authors, keywords, etc.). So-called multi-value attributes can be used to aggregate nodes just like single-value attributes. Users create these multi-value attributes by combining values of an attribute from adjacent nodes. For example, papers from Figure 3 can be aggregated by their author’s last name. Since a node can be connected to multiple nodes (e.g., paper P1 is connected to multiple authors), aggregates for the multi-value attributes may not be disjoint. For example, P1 would be in the aggregates for A1, A2, and A3. To indicate these attributes are *derived* from another node type, we add a prefix for its source to the name (e.g., Author.Lastname).

Familiar Charts

To make it easier for a broad range of people to perform visual analysis of networks, GraphTrail uses familiar charts such as bar charts, tag clouds, tables, and matrices for visualizing network item aggregations. Each chart allows people to customize the representation for the subset being viewed. Users can select the appropriate chart type, aggregation method for the network items, sizing and sorting options, and other presentation settings through standard widgets such as combo boxes (Figure 2, top). We augment traditional bar charts by with an arc diagram to see relationships between node aggregates; node aggregates are bars and the edge aggregates linking them are arcs. E.g., in Figure 2, papers in the InfoVis topic are grouped by author names and sorted by the number of papers written by each author. Bar height encodes the count of papers in it and aggregate edges linking bars show citations, with counter-clockwise directionality (i.e., papers in a left bar cite papers in a right bar via a bottom arc). Edge thickness encodes the count of citations of one author’s papers by another’s.

Three Actions for Data Exploration

GraphTrail provides three actions for exploring network data: 1) filtering and merging, 2) pivoting, and 3) cloning.

<u>Filtering and Merging:</u> Filtering is one of the most frequent and powerful exploration mechanisms. GraphTrail allows users to easily select subsets of the network and drill down into them. Inspired by Visage [26], we use a drag-and-drop metaphor to filter. Users can select subsets and drag-and-drop the subset onto the canvas, creating a chart with only that subset. Users can merge subsets by dragging them from one or more visualizations into an existing chart. Likewise, dragging edge aggregates pulls all the attached nodes. For example, one could drag an edge aggregate from Figure 2 to produce a matrix of all citations between the two author bars, showing individual paper or topic relationships.

<u>Pivoting to Connected Node Types:</u> Transitioning between different node types plays an important role in analyzing heterogeneous networks [16,19]. GraphTrail uses a many-

to-many pivot operation that moves from a subset of nodes to all linked nodes along an edge type. E.g., users can pivot from a subset of papers (all InfoVis papers) to authors linked by the "author" edge type, making a new chart with a subset of authors (all authors of those InfoVis papers). Users can then pivot on papers once again to show all the papers (not just InfoVis papers) published by those authors.

Users can also pivot within a node type. For example, users can pivot from a subset of papers along the "cites" edge type that connects papers, in either direction. The pivot operation following the edge direction would select "all papers cited by papers in a subset," while the opposite would find "all papers that cite papers in a subset." Note that this many-to-many pivot is more flexible than those provided by previous systems. FacetLens [19] supports pivots from a single node instead of a set and NetLens [16] supports pivots only between two different node types.

Data Cloning: The proper result of modifying a chart in the middle of a chain can be ambiguous. In particular, it is not clear whether we should preserve charts early in the chain for history awareness or allow changes and propagate them down the chain of resulting charts. Moreover, changing the view of a parent subset or adding/removing data can complicate the viewing of the path of filtering operations.

To limit confusion and preserve the intrinsic property of GraphTrail to provide history awareness, we lock the data and representation of any chart with dependent child charts. However, we support cloning of a chart to let users pursue alternative exploration branches from any point in a chain. The clone contains the same data as the original, but lets users edit its aggregation, representation, and even add data.

Visual Exploration History

The interaction process of GraphTrail automatically creates a visual exploration history. Showing this history helps users understand the actions that led to a visualization, recall the exploration history, and share analyses with others. When users create multiple charts via successive actions, GraphTrail connects those charts by links indicating the chain of actions. During an exploration, users leave a trail of visual and textual breadcrumbs that represents their exploration path without additional annotation effort. When zoomed out, the GraphTrail canvas represents an exploration overview. User provided annotations can help describe the "why" of an exploration (e.g., [30,11]), however, our evaluations show that annotations are not needed to describe the "how" of an analysis (e.g., the filters applied and pivots used) or for other users to infer much of the analyst's thought process.

The history of these actions is shown using parent links. Parent links are colored links that connect charts and show the action applied to create the subsequent chart; filtering is shown in gray, pivoting in red, and cloning in blue. A parent link can be moused over to highlight the entire chain of actions used to get to the subset in the chart, all the way to a root chart showing all nodes of one type. Each parent link in the path is highlighted, along with the source subset(s) in each chart in the path. This is shown using dark green in Figures 1, 4, and 5. Highlighting the entire chart (Figure 4, ROOT and GT3) indicates that its entire subset was used for a pivot (red parent link). This helps users see the history of all actions without parsing a textual query or list, though we do show that text in each chart and in parent link tooltips. The query is composed of equality statements and logical operators. For example, Figure 2 has the query: "topic = InfoVis," meaning papers with the InfoVis topic. The text query gives users an additional history view when zoomed in on a chart, though the query can become long.

As described in the previous section, data can be dragged from any chart in the exploration path to allow users to merge the results of multiple chains together or to create offshoot analyses. Thus, the history of a chart can be a directed acyclic graph in contrast to standard linear and hierarchical approaches. This is seen in Figure 4, where the MERGED chart is the union of the data in GT5 and MS5.

FIELD STUDY WITH ARCHEOLOGISTS

In this section, we report a field study with six archeologists who used GraphTrail to explore their data for three months.

Participants and Data

During the past seven years, a team of archeologists at the University of Southampton sought to answer the following research questions: "How were Iron-Age communities integrated into the political and economic structure of the Roman Empire?" and "How were urban social hierarchies within the Roman provinces structured and articulated?" To answer these questions, they collected data over the course of three years on thousands of artifacts excavated from about two hundred archeological sites in southern Spain. This dataset forms a network and consists of 12,838 nodes from 24 node types, linked by 18,927 edges with 35 types.

In the past year, six archeologists have attempted to analyze this data and understand how sites are related through their excavated artifacts. Three of them actively participated in the data analysis using GraphTrail: John, a professor and researcher in archeology for over 15 years with little computer expertise, Bob, a senior lecturer in archeological computing, and Tim, their co-advised Ph.D. student. Bob and Tim are less experienced with the data but have more expertise with analysis software. We communicated with them sporadically via email and conducted longer interviews with Tim, our main point of contact. Interviews were conducted via conference calls after each of three of their day-long analysis sessions.

Current Practice

Prior to GraphTrail, the archeology team primarily used Cytoscape [27], a visual exploration tool based on node-link diagrams. While they investigated the use of Pajek [4], they decided not to use it due to its steep learning curve.

Our interviews revealed an arduous, halting exploration process. The varied types of artifacts and the many possible derived ties between sites made it difficult to analyze the entire network in Cytoscape. Instead, they opted to generate different networks for each type of artifact and its attributes. For example, they had separate networks for ceramics of each era or for different usage. This process led them to create a large number of files. While this data processing labor was nothing compared to the tasks of data gathering, cleaning and editing, the time required made it difficult to pursue hypotheses on the spot. During collaborative data analysis sessions, further analysis was often postponed to allow one of them to generate the appropriate networks.

Results

First Impressions

We provided the archaeologists with GraphTrail a day before one of their data analysis meetings. For lack of time, we could only send the binaries and a few minutes of video demonstration. With such minimal training, we were pleasantly surprised by their report that they used GraphTrail for five hours during their full day meeting. GraphTrail was used alongside Cytoscape and Bob noted how GraphTrail *"brought them closer to the data."* They also said that every person of the team could use it, even those less familiar with databases or analysis software.

Exploration and Findings

After three months, our participants sent us a 60-page report containing a summary of their current findings. While many findings came from their previous analyses with other tools, we learned how they used GraphTrail and began to assess its strengths and weaknesses.

Tim reported that they initially used GraphTrail to *"explore the fragmentary nature of the data and generate abstract qualified groups of data."* He said that they used GraphTrail to perform queries on the fly and generate many groups of sites or artifacts with specific parameters. They created many charts in this process, and used the canvas to preserve them and remind themselves which analyses were important rather than using deep exploration paths. Bob mentioned that GraphTrail helped them browse the data differently from other tools and stated multiple times how *"quickly"* they could perform queries. From our discussions with the team, it appears that GraphTrail provided a more systematic analysis through the creation of multiple high level overviews of the data through various aggregations.

Tim stated that GraphTrail *"confronted them with the data"* by enabling them to rapidly explore different aggregations of sites and artifacts. GraphTrail also helped study patterns found previously using Cytoscape. For example, they had uncovered two uncommon types of ceramics in Iberian sites and thought that their appearance may be significant. By investigating this pattern further using GraphTrail, they discovered that it was due to data error. In GraphTrail they were able to group this data as one entity for further study without having to go back to the database or change files. They also made serendipitous discoveries. For example, they *"... noticed that some forms* [of ceramics] *were only found in distinct groups of sites, and that these sites had no other ceramic forms in their assemblages."* These findings led to new questions and directions for further exploration.

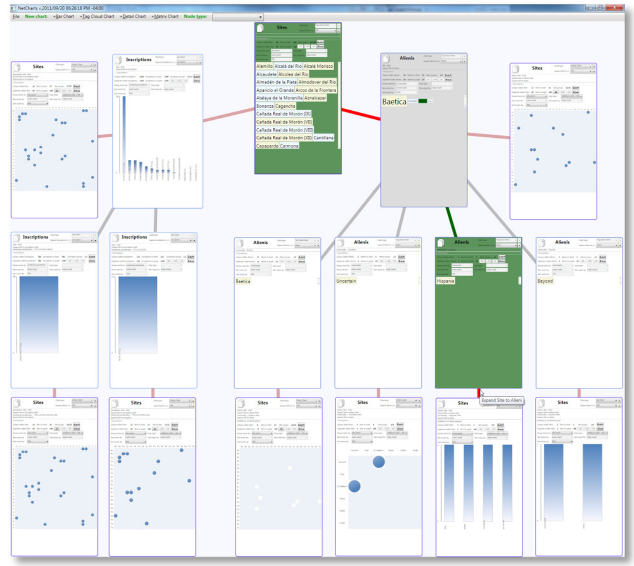

Figure 5. The dual-path exploration Tim used for the archaeology network. The right branch shows the initial exploration of sites grouped by alieni, with detail for each. The left branch shows sites grouped by inscription, with detail charts for the two most common inscriptions.

Exposing the Exploration Process

GraphTrail differs from other analysis tools in that it exposes the exploration process to its user. When going over one analysis session with Tim, we observed how the exposition of his exploration triggered more questions and pushed the analysis further. Tim started by exploring how sites are connected depending on the origin of foreign individuals (alieni) found in them. *"For the Beyond category there's nothing, you see, no real links and for Hispania there are no real links here either. I also did that for Baetica and there are links here, so it is interesting […] It is interesting to pick these things that are different, anomalies."* This finding led him to reflect on his exploration and to a desire to apply the same process to another part of the data. *"I want to see, at this point, what visualizations like this can tell me if I do exactly the same thing for the inscriptions."* He then conducted the same series of operations with the inscriptions artifact. Figure 5 shows this dual-path exploration. On the right is an initial analysis of sites grouped by alieni, and on the left are sites grouped by inscriptions. The matrix charts were used to support a discussion on the correlation of these artifacts.

Exploration History

To assess how well GraphTrail supports exploration histories, we asked Tim to present to us an analysis session that he had completed the previous week. We also asked him to describe an exploration he was not aware of, performed by one of his colleagues.

Tim mostly used the spatial organization of his analysis to explain his past exploration (Figure 5). He pointed to the top-central chart as the origin of his search and pointed to the two branches to describe them at a high level. Then, he interacted with the tool and navigated the canvas as he explained his analysis in more detail. He zoomed in on each of the charts moving from left to right as he described the parallel operations. It is interesting to note that he used the content of the charts as a reminder of what the chart was and did not rely on the visual history of his operations provided on mouse over. After the session, he claimed that he did not need this feedback as he could still recall what he was trying to achieve at a high level and thus could easily deduce the operations he had performed to reach his goal.

In contrast, when he described his colleague's analysis he could not rely on the spatial organization clues, stating that, "*According to my visual logic, I would start at the center*" and later, after studying the canvas, "*this* [spatial organization] *is irrational, it does not make sense at all!*" Then he reflected on the fact that it was a personal choice: *"the* way ***I*** *structured it makes sense to* ***me****."* To make sense of his colleague's exploration, he started using GraphTrail's visual feedback. With this feedback, the contents of the charts, and their associated queries, he quickly managed to make sense of the analysis. He was quite surprised that his colleague ended up finding the same correlation between alieni and inscriptions but with a totally different approach. Instead of parallel explorations to compare sites aggregated by alieni origin or inscriptions, his colleague investigated two particular sites and their alieni origin and inscriptions. Surprised by the findings from this "*simple query*" approach, Tim commented that the colleague may be more apt at this type of exploration.

QUALITATIVE LAB STUDY WITH HCI RESEARCHERS

We conducted a study in a laboratory setting with three goals: 1) to evaluate if novice users could make findings with GraphTrail, 2) to investigate how they would use the visual history mechanism, and 3) to find any usability issues in order to improve the system.

Participants and Tasks

We recruited ten researchers and interns (eight males, two females) who were familiar with the field of HCI from our institution and asked them to analyze the CHI paper data. Each session began with a hands-on analysis demonstrating all features of the system, followed by two testing phases. In the first phase, participants used GraphTrail to explore the data. We prompted participants with a high-level question ("What can you tell me about HCI research in Georgia?") and asked them to articulate their intentions, actions, problems, and any findings (i.e., facts in the data unknown by the participant). Example findings are that HCI research in Georgia is dominated by Georgia Tech's group that focuses on CSCW and Multi-modal UI. In the second phase, to assess if participants could interpret analyses created by others, we gave participants two pre-created explorations from a different part of the dataset and asked them to describe any findings and their reasoning process. As participants successfully followed the analyses, we raised the complexity. The most difficult analysis included 18 charts with two comparison tasks and one detail task.

Results

Phase 1: Exploration

In roughly 30 minutes, all participants discovered all the findings shown in Figure 4's GT charts (the findings reported by NetLens [16] for this dataset). They were able to identify the correct affiliation (including name variants) and its main researchers. They could assess their patterns of collaboration and identify key paper topics and publication rhythm over time. Some participants went further in their analysis and linked patterns in the data to events they were familiar with (e.g., Elizabeth Mynatt joining Georgia Tech).

We collated published findings in this dataset by reviewing the case studies and examples in three other papers that used it: PaperLens [18], NetLens [16], and FacetLens [19]. GraphTrail supported all of these analyses, excluding some for PaperLens that require dataset-specific computations we do not support. Some participants made findings that were not reported by or could not possibly be discovered with other systems due to their design. One participant noted that both Elizabeth Mynatt and Jonathan Grudin (not from Georgia Tech) self-cite rarely given their publication volume. Another example is finding all papers written by all Japanese authors; an easy task in GraphTrail which is impossible in FacetLens since it pivots from only one item.

During this phase, we learned that most participants struggled with the notion of pivoting. It took most participants approximately half the session to fully grasp the operation and its result.

Phase 2: Exploration Histories

We hypothesized that this task would be difficult, even for a small analysis, because participants did not know the goal of the analysis and might not have used the same operations themselves to achieve the goal. Each participant was able to follow the expert analyses we provided without much difficulty, even after extra analysis branches were added. Three (out of ten) made all findings expected while the others made approximately 90% of the findings (the missing 10% dealt with the citation pivoting that GraphTrail did not directly support at the time).

Four participants claimed the *spatial organization* of charts provided a useful overview, aiding in understanding branches and related charts. Participants also found *parent-link highlighting* and *tooltips* helpful for reading analyses, using them to correct misinterpretations and to understand large chunks of an analysis at once when zoomed out.

DISCUSSION AND LIMITATIONS

More Chart Types and Interactions

We learned through the field study that Cytoscape's node-link diagrams were often used by our participants to report their findings. On multiple occasions, they would spot an interesting link pattern using the hybrid bar chart, analyze it further in a matrix, and present it in their report using a node-link diagram. The archeologists also performed many comparisons between charts, which is currently difficult because the sizes of aggregates are relative to each chart. Providing normalized scales across charts may aid direct comparisons. Also, brushing and linking charts would show related data and improve comparisons. We are considering adding chart types better suited to comparisons such as stacked bar charts and ranked lists, as well as task-specific charts like maps for the archaeologists.

Spatial Memory vs. Management Cost

Inspired by Data Mountain's use of users' spatial memory [25] for recalling their explorations, we let users place each chart they create. While spatial organization requires user management, we observed during the field study that it is tightly-coupled with the exploration process. Users placed each chart at an appropriate location right after it was created. Hence, we believe that the memory and cognitive benefit outweighs the cost. In our studies most participants reported that it was not much of an extra burden to arrange the charts as they were created. The ability to create new charts with a drag-and-drop interaction was acknowledged by one participant as the "*extreme flexibility offered by the system.*" We observed that the spatial organization helped participants organize and remember their process.

Scalability

There are two levels of scalability to consider in GraphTrail: 1) the number and types of nodes and edges visualized and 2) the number of charts in the exploration history. Visualizing a large multivariate and heterogeneous network is challenging, but our approach of using familiar charts and aggregating network items based on their attributes has proven to be effective at reducing the complexity. The data used for our studies are sufficiently large to be challenging to display using node-link diagrams (CHI data with ~10K nodes and ~20K edges and Archaeology data having ~13K nodes and ~20K edges). Moreover, examining the 35 types of relationships between the archaeology data's 24 node types is beyond the capabilities of most analysis tools, though handled well by the GraphTrail many-to-many pivot.

Our participants usually created analyses of 20-30 charts at most, starting a new file for a new exploration. When asked why, they said it made sense to start a new file for each new question. While our participants did not mention scalability as a shortcoming, explorations over multiple sessions/users may become very large. GraphTrail does not limit how many charts users can create, but having a large exploration history introduces a layout management cost for the user. One solution to ease this management is to reduce the number of charts and compress paths to "landmark" results, throwing away intermediate steps [20,14]. However, when this approach was suggested to users, they argued strongly to keep the intermediate charts. Therefore, collapsible branches and semantic or continuously variable zoom [8] would likely be the most effective simplification technique that would retain intermediate charts. Also, parent charts can be closed [26] or reused for propagating parameter changes. While the number of charts may be reduced, the history of operations becomes more difficult to track (as older charts are collapsed or modified). By preventing these kinds of changes to parent charts, GraphTrail reduces ambiguity and show the history of all explorations.

Exploration History Model

The analysis history we keep and display to the user is not the complete history of their actions – instead it is a subset of meaningful actions necessary to understand the displayed analysis. Showing each visual transformation, individual mouse events, or an explicit timeline of user actions could be overwhelming to users and of debatable usefulness for recall and sharing. However, this detailed information may be necessary for user behavior analyses and understanding how to present it effectively is an open question.

CONCLUSION

This paper presents GraphTrail, an interactive visualization system for analyzing large multivariate and heterogeneous networks while displaying exploration history. GraphTrail uses familiar charts to show aggregations of nodes and edges, and presents them on a canvas where users can employ drag-and-drop operations to filter and create charts. A pivot mechanism lets users move between node types along linking edges. Operations leave visual breadcrumbs of a user's exploration by creating parent links and textual cues, documenting the analyst's reasoning process.

We present results from the long-term (3-month) use of GraphTrail by a team of archeologists as well as a qualitative lab study with ten novices. Our participants responded positively to the familiar charts and drag-and-drop filtering methods, and both archeologists and novices were able to make findings in large multivariate and heterogeneous networks. We observed the benefit of exposing exploration history to users: GraphTrail inspired further exploration and users could recall their findings and the exact exploration process to arrive at them. Both studies showed that users could interpret explorations of others using visual history feedback. In the future, we plan to add new chart types and continue our field study with the archeologists.

REFERENCES

1. Abello, J. and van Ham, F. Matrix Zoom: a visual interface to semi-external graphs. *Proc. InfoVis '04*, (2004), 183-190.

2. Abello, J., van Ham, F., and Krishnan, N. ASK-GraphView: a large scale graph visualization system. *IEEE TVCG (InfoVis '06) 12*, 5 (2006), 669-676.
3. Adar, E. GUESS: a language and interface for graph exploration. *Proc. CHI '06*, (2006), 791-800.
4. Batagelj, V. and Mrvar, A. Pajek – Program for large network analysis. *Connections 21*, (1998), 47-57.
5. Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T., and Vo, H.T. VisTrails: enabling interactive multiple-view visualizations. *Proc. VIS '05*, (2005), 135-142.
6. Bezerianos, A., Chevalier, F, Dragicevic, P., Elmqvist, N., and Fekete, J.-D. GraphDice: a system for exploring multivariate social networks. *Computer Graphics Forum 29*, 3 (2010), 863-872.
7. Blaas, J., Botha, C., Grundy, E., Jones, M., Laramee, R., and Post, F. Smooth graphs for visual exploration of higher-order state transitions. *IEEE TVCG (InfoVis '09) 15*, 6 (2009), 969-976.
8. Dill, J., Bartram, L., Ho, A., and Henigman, F.A continuously variable zoom for navigating large hierarchical networks. *Proc. SMC '94*, (1994), 386-390.
9. Fekete, J.-D., Wang, D., Dang, N., Aris, A., and Plaisant, C. Overlaying graph links on treemaps. *Posters compendium of InfoVis '03*, (2003), 82-83.
10. Freire, M., Plaisant, C., Shneiderman, B., and Golbeck, J. ManyNets: an interface for multiple network analysis and visualization. *Proc. CHI '10*, (2010), 213-222.
11. Heer, J., Mackinlay, J.D., Stolte, C., and Agrawala, M. Graphical histories for visualization: supporting analysis, communication, and evaluation. *IEEE TVCG (InfoVis '08) 14*, 6 (2008), 1189-1196.
12. Henry, N. and Fekete, J.-D. MatrixExplorer: a dual-representation system to explore social networks. *IEEE TVCG (InfoVis '06) 12*, 5 (2006), 677-684.
13. Henry, N., Fekete, J.-D., and McGuffin, M.J. NodeTrix: a hybrid visualization of social networks. *IEEE TVCG (InfoVis '07) 13*, 6 (2007), 1302-1309.
14. Jankun-Kelly, T.J., Ma, K.-L., and Gertz, M. A model and framework for visualization exploration. *IEEE TVCG 13*, 2 (2007), 357-369.
15. Kadivar, N., Chen, V., Dunsmuir, D., Lee, E., Qian, C., Dill, J., Shaw, C., and Woodbury, R. Capturing and supporting the analysis process. *Proc. VAST '09*, (2009), 131-138.
16. Kang, H., Plaisant, C., Lee, B., and Bederson, B. NetLens: Iterative exploration of content-actor network data. *Information Visualization 6*, (2007), 18-31.
17. Kang, Y-a. and Stasko, J. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. *Proc. VAST '11*, (2011), 21-30.
18. Lee, B., Czerwinski, M., Robertson, G., and Bederson, B. Understanding research trends in conferences using PaperLens. *Ext. Abstracts. CHI '05*, (2005), 1969-1972.
19. Lee, B., Smith, G., Robertson, G.G., Czerwinski, M., and Tan, D.S. FacetLens: exposing trends and relationships to support sensemaking within faceted datasets. *Proc. CHI '09*, (2009), 1293-1302.
20. Lee, J.P. and Grinstein, G. An architecture for retaining and analyzing visual explorations of databases. *Proc. Vis '95*, (1995), 101-108.
21. Lipford, H.R., Stuke, F., Dou, W., Hawkins, M.E., and Chang, R. Helping users recall their reasoning process. *Proc. VAST '10,* (2010), 187-194.
22. North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B., and Fink, G. Analytic provenance: process + interaction + insight. *Ext. Abst. CHI '11*, (2011), 33-36.
23. Perer, A. and Shneiderman, B. Balancing systematic and flexible exploration of social networks. *IEEE TVCG (InfoVis '06) 12*, 5 (2006), 693-700.
24. Pretorius, A.J. and van Wijk, J.J. Visual analysis of multivariate state transition graphs. *IEEE TVCG (InfoVis '06) 12*, 5 (2006), 685-692.
25. Robertson, G., Czerwinski, M., Larson, K., Robbins, D.C., Thiel, D., and van Dantzich, M. Data Mountain: using spatial memory for document management. *Proc. UIST '98*, (1998), 153-162.
26. Roth, S.F., Lucas, P., Senn, J.A., Gomberg, C.C., Burks, M.B., Stroffolino, P.J., Kolojejchick, J.A., and Dunmire, C. Visage: a user interface environment for exploring information. *Proc. InfoVis '96*, (1996), 3-12.
27. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research 13*, 11 (2003), 2498-2504.
28. Shneiderman, B. and Aris, A. Network visualization by semantic substrates. *IEEE TVCG (InfoVis '06) 12*, 5 (2006), 733-740.
29. Shrinivasan, Y.B. and van Wijk, J.J. Supporting the analytical reasoning process in information visualization. *Proc. CHI '08*, (2008), 1237-1246.
30. Stasko, J., Gorg, C., Liu, Z., and Singhal, K. Jigsaw: supporting investigative analysis through interactive visualization. *Proc. VAST '07*, (2007), 131-138.
31. Ware, C., Gilman, A.T., and Bobrow, R.J. Visual thinking with an interactive diagram. *Proc. Diagrams '08*, (2008), 118-126.
32. Wattenberg, M. Visual exploration of multivariate graphs. *Proc. CHI '06*, (2006), 811-819.
33. Zhao, S., Chignell, M.H., and McGuffin, M.J. Elastic hierarchies: combining treemaps and node-link diagrams. *Proc. InfoVis '05*, (2005), 57-64.