

Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes

Aline Menin, Ricardo Cava, Carla Maria Dal Sasso Freitas, Olivier Corby, Marco Winckler

▶ **To cite this version:**

Aline Menin, Ricardo Cava, Carla Maria Dal Sasso Freitas, Olivier Corby, Marco Winckler. Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes. IV 2021 - 25th International Conference Information Visualisation, Jul 2021, Melbourne / Virtual, Australia. pp.21-28, 10.1109/IV53921.2021.00014 . hal-03292172

HAL Id: hal-03292172
https://hal.science/hal-03292172

Submitted on 20 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes

Aline Menin
I3S (UMR 7271)
University Côte d'Azur, Inria, CNRS
Sophia Antipolis, France
aline.menin@inria.fr

Ricardo Cava
Sul-rio-grandense Federal Institute
Pelotas, Brazil
ricard.cava@gmail.com

Carla Maria Dal Sasso Freitas
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
carla@inf.ufrgs.br

Olivier Corby
I3S (UMR 7271)
University Côte d'Azur, Inria, CNRS
Sophia Antipolis, France
olivier.corby@inria.fr

Marco Winckler
I3S (UMR 7271)
University Côte d'Azur, Inria, CNRS
Sophia Antipolis, France
marco.winckler@inria.fr

***Abstract*—Visualization techniques are useful tools to explore data by enabling the discovery of meaningful patterns and causal relationships. The discovery process is often exploratory and requires multiple views to support analyzing different or complementary perspectives to the data. In this context, analytic provenance shows great potential to understand users' reasoning process through the study of their interactions on multiple view systems. In this paper, we present an approach based on the concept of chained views to support the incremental exploration of large, multidimensional datasets. Our goal is to provide visual representation of provenance information to enable users to retrace their analytical actions and to discover alternative exploratory paths without loosing information on previous analyses. We demonstrate that our implementation of the approach, MGExplorer (Multidimensional Graph Explorer), allows users to explore different perspectives to a dataset by modifying the input graph topology, choosing visualization techniques, arranging the visualization space in meaningful ways to the ongoing analysis and retracing their analytical actions. MGExplorer combines multiple visualization techniques and visual querying while representing provenance information as segments connecting views, which each supports selection operations that help define subsets of the current dataset to be explored by a different view. We demonstrate the usage of the tool through a study case where we explore co-authorship data. We assess the approach through performance metrics, temporal ordering of tasks, number of physical actions, and amount of information to be recalled in-between actions applied to the chosen visual exploration scenarios using chained views.**

***Index Terms*—Analytical provenance, Multiple views, Chained views, Multidimensional data exploration, Provenance visualization**

I. Introduction

Visual analytics is widely known for facilitating human reasoning through interactive tools embedding visual representations that highlight and reveal the relationships within data. It is a suitable approach to support decision-making processes in application domains as diverse as public health [1], social media [2], and finance [3], where professionals are confronted with the analysis of huge datasets, often characterized by multiple attributes or dimensions. Nonetheless, as the number data dimensions increases, using a single view to display as much information as possible at once might minimize the need for exploration but it can engender cognitive overload on users and create visual clutter-related issues [4].

The coordinated multiple views (CMVs) paradigm tackles the problem of representing multidimensional data via a predefined set of views [5] that represent different perspectives to the data, while providing coordinating operations between views to support reasoning [6]. However, the great variety of available visualization techniques to explore multidimensional datasets [7]–[10] makes it unfeasible to determine a single combination of visualization techniques capable of solving every domain-related task. Moreover, most CMV-based systems usually deal with single datasets, being of little help when user reasoning requires comparing different data subsets.

Typically, in an exploratory context, the user has no defined goal and is looking for no particular outcome [11]. Though, when finding something interesting, users should be able to retrace their exploratory path to explain how they found the results. Moreover, the validation of hypotheses might require branching out the exploratory path to compare data observed in different views. Evidence of these exploratory processes has promoted rapid growth in research on analytical provenance [12][13], including techniques to capture, visualize, and analyze provenance information. Most visualization systems record the user's actions through history graphs [14], which is often not enough to analyze the analytical process [15].

Contributions. The primary contribution of this paper is a flexible visualization approach based on the concept of chained views, capable of depicting analytical provenance via a sequence of views, while supporting one or more visualization techniques applied to one or more datasets. Thus, it supports visual analysis via multiple alternative exploration scenarios that can be retraced and modified. We compare our solution with other existing approaches, highlighting the need

for innovative and flexible tools for exploring large datasets, and demonstrate the feasibility of the chained views approach through an interactive visual tool called MGExplorer, created to assist the exploration of multivariate networks.

The remainder of this paper is organized as follows. Section II summarizes related works on multiple and integrated views and analytical provenance visualization. Section III presents MGExplorer, a visualization interface to assist the exploration of multidimensional networks. Section IV demonstrates our approach through a use case employing a co-authorship dataset. Section V presents an evaluation of the effectiveness and efficiency of MGExplorer. Section VI discusses the contributions of this work and concludes the paper.

II. Related Work

Several works in the literature use chained views to explore large, multidimensional datasets due to its ability to provide users with means of (i) choosing the visualization techniques and data subsets meaningful to their analyses, (ii) retracing their path to understand how they arrived at a certain point in the exploration and to identify the starting points of specific exploration paths, and (iii) creating alternative paths from previous points in the exploratory process.

A. Chained Views to Explore Multidimensional Datasets

Table I summarizes the visualization tools surveyed in this section according to the function of the proposed visualization techniques (e.g., identifying variables' distribution, comparing data items); whether they allow the user to modify the topology of the input data to answer different questions; whether they provide support for an interactive history of visualizations created over the exploration process, allowing the user to explore and view provenance information even when visualizations are not displayed on the screen; and the purpose of authors for using chained views, whether it is meant for analytical provenance studies or a different motivation.

TABLE I
COMPARISON OF FEATURES FOUND IN THE SURVEYED WORKS: VISUALIZATION TYPE (**D**: DISTRIBUTION, **T**: TEMPORAL, **R**: RELATIONSHIP, **C**: COMPARISON, **F**: FLOW, AND **CL**: CLUSTERING); SUPPORT FOR TOPOLOGY CHANGE; INTERACTIVE HISTORY; PURPOSE OF USING CHAINED VIEWS.

Tool	Visualization	Topology Change	Interactive History	Purpose of Chained Views
Connected Charts	D, R, C	No	No	Retrace history of analytical steps
Domino	D, R, C, F	Yes	No	Represent the strength of relationship between charts
GraphTrail	D, R, C	Yes	No	Retrace history of analytical steps
SOMFlow	T, R	No	No	Analytical provenance
MGExplorer	T, R, C, CL	Yes	Yes	Analytical provenance

The ConnectedCharts [16] tool supports instantiating and placing different charts freely on the display area; curves connecting charts show the correspondence between data elements or axes. Although the visualization anchors the curves to the edges of the charts and draws only the curves created by the user to avoid occlusion, it is still overlaid by several lines connecting the multiple geometric shapes in each chart, which could compromise readability. Moreover, the relationships between views are only item-based, while MGExplorer also supports relationships based on subsets of data.

Domino [17] allows the user to arrange, combine, and manipulate subsets of data, while explicitly representing the relationships between those subsets and allowing the user to choose suitable visualization techniques. However, it uses a juxtaposed layout that restrain the possible exploratory paths by placing views side-by-side and may cause ambiguities by juxtaposing non-semantically related paths.

GraphTrail [18] and SOMFlow [19] systems support the exploration of large datasets by using multiple connected views and diverse visualization techniques. GraphTrail supports the exploration of large multivariate, heterogeneous networks through drag-and-drop interactions used to refine subsets of data in a new view while showing users' exploration history by lines connecting the views. In SOMFlow, each view shows a cluster refinement of a dataset and the links represent the analytical workflow of the exploration partition process. Both approaches are restricted to subsets of a unique dataset, while MGExplorer also allows instantiating views using data from external datasets via querying operations.

It is worthy to notice that, although these four systems allow the user to follow alternative exploratory paths by hiding the current views and instantiating new ones, they fail to provide an alternative visual presentation of provenance information that allow users to compare or recover previous exploratory paths. For that purpose, our approach includes both the visual connection between views and an interactive history panel, displaying and allowing interaction with provenance information during the whole experience.

B. Provenance Data Visualization

Analytical provenance focuses on understanding users' reasoning process by studying their interactions while using a visualization system [20]. In this work, we are interested in *visualizing* provenance information to allow the users to interact with it to understand their exploratory paths and support further studies of these data to improve user experience. In this section, we discuss provenance visualization according to three aspects of analytic provenance: (i) *why* to analyze provenance data, (ii) *what* are the types of provenance data and ways to encode it, and (iii) *how* to analyze provenance data.

There are several reasons (*why*) for provenance analysis, such as understanding users, evaluating systems and algorithms, building adaptive systems, model steering, replicating, verifying and re-analyzing, reporting and storytelling [13]. Our work aims to allow users to recreate their analytical reasoning process, while supporting verification, replication, or reapplication of analysis sessions. This could be supported by history trees containing enough metadata to allow the user to search and retrieve visualizations [21] through features such

as the ones provided by VisTrails [22], which allows the user to create, change, and compare visualizations by exploring the graphical representation of the dataflows. The links between visualizations in GraphTrail [18] and SOMFlow [19] represent the workflow, allowing users to understand the actions performed from a visualization or analytical step to another.

The second aspect (*what*) refers to the way provenance data is represented. Among the schemes of provenance data representation (e.g., grammars, models, graphs), the sequence-based scheme is the most common [13], and represents the user's interactions as a list of actions. Graphs are used to connect entities (or concepts) that change state during the exploration process, or temporal events, as in history graphs (e.g., VisTrails, GraphTrail, and SOMFlow). Further to allowing the user to go back and forth in the exploration path, graph encoding can, for instance, allow a user to interact directly with the history (provenance) graph to generate a story of the user's analysis [23]. MGExplorer is based on a history graph, which allows forking alternative exploratory paths or changes in parameters of a previously created view; all the views created by users during an exploratory process can be shown or hidden and used to record their flow.

Finally, as for the *how* aspect, previous works have addressed provenance analysis using methods such as classification and statistical modeling techniques, pattern analysis, probabilistic techniques for prediction, grammar and scripts analysis, and interactive visual analysis [13]. Although this paper do not target the analysis of provenance data, the explicit representation of provenance information supported by our approach allows anyone to perform an interactive visual analysis on the exploratory flows by querying the parameters used in each step of the process.

III. MGExplorer

A. Design rationale

Exploring large datasets through multiple views systems can lead to multiple exploratory paths with arbitrary starting points. Thus, to assess users' reasoning during the exploration process, one should be able to retrace the dependency between the views. For that, we developed MGExplorer (**M**ultidimensional **G**raph **Explorer**), a visualization tool to assist the exploration of multidimensional network data. The tool uses the chained views approach to allows users to combine multiple visualization techniques during the exploration process while recording descriptive information of the exploration path, thus supporting analytical provenance studies.

Mainly, MGExplorer provides visualization techniques that support graph data exploration by revealing different types of relationships within the dataset (e.g., network, clusters, pairwise, and order). The tool can encode different data dimensions by modifying the graph topology. To avoid cluttering problems, the visualization layout (i.e., the arrangement of views) is flexible, allowing users to freely position the views, hide and show views of interest without losing the connections between them. It also supports interaction with the history of analytical actions performed during the exploration of the dataset. Furthermore, to ensure accessibility, MGExplorer is based on web technologies, i.e., JavaScript, which D3 (Data-Driven Documents) library is particularly used in front-ends, and the nodejs library manages the access to SPARQL endpoints, facilitating the retrieval of different datasets for exploration[1]. The framework is generic enough to support the exploration of data from different sources (e.g., linked data [24], web services, and databases), as it includes in the server side a data transformation process that adapts the input dataset to the system's input format. Thus, data visualization through MGExplorer comprises mainly two complex processes: (1) data treatment and (2) data exploration.

B. Overview of the Exploratory Process

We use MGExplorer to explore a dataset describing a co-authorship network, i.e., a graph where authors are explicitly related due to the publications they have together. The data exploration process unfolds into two phases (Fig. 1): (1) the *overview* phase, which is based on visualizing the network through a node-edge diagram. This visualization allows the user to understand the co-authorship clusters within the data; and (2) the *exploratory* phase, which enables users to select items of interest in the node-edge diagram, subsetting the data to explore it through different, complementary visualization techniques. This way, the generic aspect of the MGExplorer framework enables the combination of multiple visualizations to allow both the comparison of two or more different subsets of data through a particular perspective generated by a particular visualization and the comparison of multiple representations of the same subset of data using multiple, complementary visualization techniques.

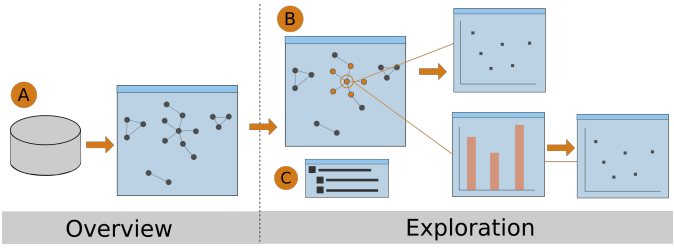


Fig. 1. Overview of MGExplorer framework. (a) The NodeEdge diagram provides an overview of the dataset. (b) Filtering operations enable further exploration of items/subsets of interest through different visualization techniques. (c) A history panel records users' actions during the exploration.

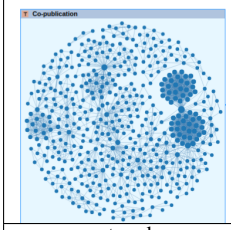
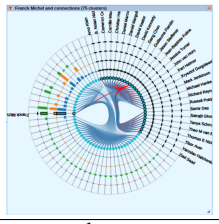
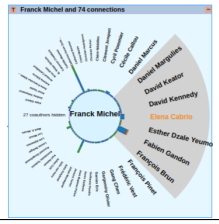
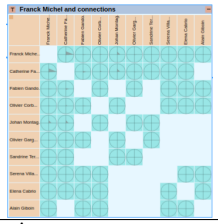
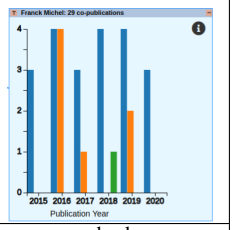
C. Visualization Techniques

Table II presents the set of five visualization techniques currently available in MGExplorer to support data exploration.

The **NodeEdge** diagram shows nodes as items and edges between them as relationships. It provides an overview of any network defined within the dataset according to different criteria (e.g., keywords, co-publications, etc.).

[1]The system is available at http://covid19.i3s.unice.fr:8080/

TABLE II
VISUALIZATION TECHNIQUES AVAILABLE IN MGEXPLORER AND TYPES OF RELATIONSHIP BETWEEN ITEM SETS PROVIDED.

NodeEdge	ClusterVis	IRIS	GlyphMatrix	Bar Chart
				
network	clusters	pairwise		order by

The **ClusterVis** [25] technique depicts clusters according to some relationship among data items. It has a multi-ring layout, where the innermost ring is formed by the data items (represented by circles), and the remaining rings display the data attributes (represented by rectangles). The items belonging to the same cluster are connected via curved lines.

The **IRIS** technique allows isolating a data item of interest (at the center) and showing the remaining data items with which it has a relationship in a circular view [26]. The data attributes of such pairwise relationships are encoded by the height and color of a bar placed between the item of interest and each related item. Upon clicking on an item, it is placed on the field of view, switching the IRIS' focus.

The **GlyphMatrix** [27] technique is based on a matrix where rows and columns represent data items in a cluster, and the cells contain glyphs encoding attributes that describe a pairwise relationship. The default glyph is a star-plot-like shape, with a variable number of axes used to encode values of selected data attributes. Hovering over a glyph in the matrix enlarges it, allowing to see the data attributes' details.

The **Bar Chart** technique shows the distribution of data attributes' value for an item or set of items. In our case study, the x-axis encodes temporal information, while the y-axis encodes the counting of co-publications. The data is displayed as a single bar per time period or multiple colored bars to represent categorical information of attributes.

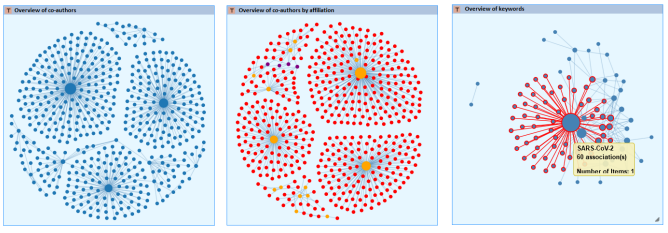


Fig. 2. Using NodeEdge technique to represent: a) a co-authorship network, b) co-authorship network, with color encoding authors' affiliation(s), and c) co-occurrence of keywords in a set of scientific publications.

Each view is a self-contained element, which includes a visualization technique and supports subsetting operations to allow further exploration of subsets of data through different views. The views can be dragged, allowing the user to rearrange the visualization space in meaningful ways to the ongoing analysis. They are connected via line segments, which reveal their dependencies and enable tracing back the exploration path, thus preserving provenance information.

MGExplorer starts displaying an overview of the data through the NodeEdge visualization and a History panel (Fig. 1-C), which displays the exploration path in a hierarchical format to indicate the dependencies between views and supports quick recovery of the multiple analytical paths that emerge from a particular view. These views are displayed throughout the whole exploration process. Recalling that in an exploratory visualization, the user would inspect the dataset without any particular goal or expected outcome, the History panel is interactive, enabling the user to revisit any of the previous visualizations and hide any of the currently displayed views. This way, the user can clean the visualization space while focusing on what is relevant to the ongoing analysis.

Figure 2 illustrates how MGExplorer supports the visualization of different perspectives to the dataset by modifying the graph topology. For the sake of simplicity we use only three views with the NodeEdge visualization to represent a dataset describing co-authorship information, where nodes represent: a) co-authors, b) co-authors, with color encoding authors' affiliation(s), and c) keywords used in the publications.

IV. USE CASE

In this use case, we demonstrate the usage of MGExplorer's visual and interactive tools to explore a co-authorship dataset extracted from the HAL (open-access repository) SPARQL Endpoint[2] through a query that retrieved every publication co-authored by at least one member of a given research organization between the years of 2015 and 2020. The query's results (2,603 RDF[3] triples describing different publications) went through a data transformation process to extract the clusters of co-authors and the descriptive information of their publications (e.g., publication year, type, etc.). The resulting graph contains 497 nodes (authors) and 2,080 edges (connections between authors). In particular, we describe the exploratory process for solving the following analytical task: "*determine the impact of recurrent co-authorship to the total number of publications of a particular author*".

[2]https://data.archives-ouvertes.fr/doc/sparql
[3]Resource Description Framework

There are mainly two indicators in the system that allow measuring the impact of recurrent co-authorship: (i) the number of co-publications and (ii) their distribution over time. For the sake of simplicity, our author of interest will be the one with the most co-authors within the dataset. To solve the task, we need to split it into smaller tasks:

1) to determine the author with the highest number of co-authors (author A);
2) to determine their most recurrent co-author (author B), i.e., the one with author A has the most co-publications; and
3) to compare the number of co-publications between authors A and B with the total number of publications co-authored by author A.

We begin the exploration by searching for the largest sub-graph in the NodeEdge view, which represents the author with the most connections. We can quickly spot two large sub-graphs connecting to the author Franck Michel (Fig. 3a), who is the author with most co-authors in this dataset. Thus, we continue the analysis with this author as our object of interest, i.e., author A. From this point, the different available visualizations support alternative scenarios to perform the task. Although the temporal distribution of co-publications can only be explored trough the Bar chart technique, identifying the most recurrent co-author can be done through any of the remaining visualizations, resulting on at least three different exploration scenarios. Therefore, we will demonstrate the versatility of our approach by performing this analytical task through two different scenarios, described hereafter.

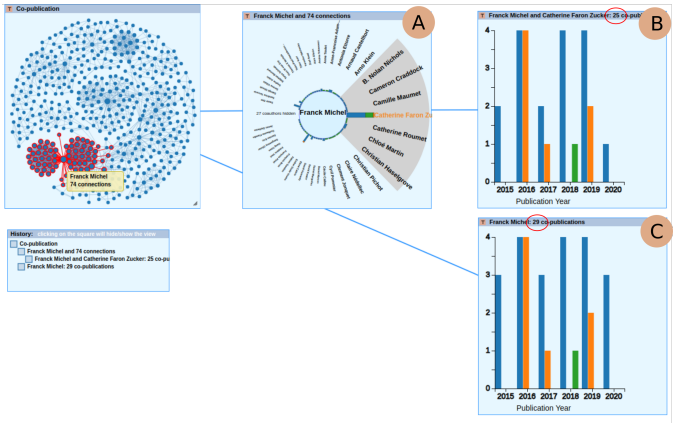


Fig. 3. Exploration path of Scenario 1.

A. Scenario 1

Figure 3 shows the exploration path of this scenario. To determine the author with whom author A has the most publications, we choose the IRIS visualization technique (Fig. 3a), which we launch by right-clicking on the node representing author A and then selecting the technique on the context menu. In the IRIS view, we search for the longest bar, which length encodes the number of co-publications between two authors. Upon clicking on the bar, the system centers it in the IRIS view, allowing us to inspect the data. We observe that Catherine Faron Zucker is author A's most recurrent co-author, whom we refer to as author B. To compare the number and distribution of their co-publications, we choose the Bar chart technique representing the temporal distribution of publications co-authored by authors A and B (Fig. 3b). We display it by right-clicking on the colored bar between the authors' names in the IRIS and then selecting the Bar chart on the context menu. Then, we revisit the NodeEdge view. There, we trigger the context menu on the node representing author A and select the Bar chart technique, which shows the temporal distribution of publications co-authored by author A (Fig. 3c). Using the information displayed on the views' titles, we observe that author A has co-authored twenty-nine publications, of which author B co-authored twenty-five. This information allows us to infer that the co-authorship between authors A and B greatly impacts the total number of publications co-authored by author A. Further, we could inspect both Bar charts to identify the importance of the collaboration over the years: the differences in the number of publications are visible in 2015, 2017, and 2020.

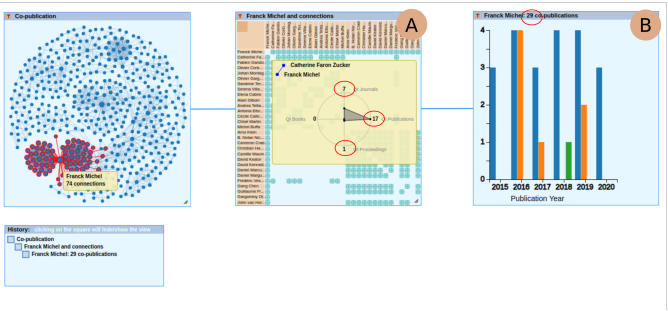


Fig. 4. Exploration path of Scenario 2.

B. Scenario 2

In this scenario, we determine author A's most recurrent co-author using the GlyphMatrix technique (Fig. 4-A), which we launched from the NodeEdge view through the context menu associated with the node representing author A. To ease the searching process, we sort the authors in the matrix in descending order of the number of co-publications, which is done by modifying the view settings on the panel available at the top-left corner of the view. This way, we know that the cell at the leftmost upper corner provides co-publication information between author A and its most recurrent co-author, Catherine Faron Zucker, hereafter called author B. Thus, we hover over the cell to inspect the glyph, which gives us the number of co-publications per category (7 articles in journals, 17 conference papers, and one article in proceedings). To compare this information with the total number of author A's co-publications, we launch the Bar chart technique via the context menu that we trigger right-clicking on the author's name displayed on the rows of the matrix. The Bar chart (Fig. 4-B) shows the temporal distribution of author A's 29 co-publications per category. We can infer the importance

of author B's co-authorship via the only four publications authored by author A that are not co-authored by author B.

V. Evaluation

This section presents a preliminary evaluation of MGExplorer using objective metrics allowing a fair comparison of alternative exploration paths, i.e., the exploratory scenarios for finding co-authors information described in Section III. We compare those with the exploratory scenario required to find co-authorship information on the Web site HAL[4], which gives a list of publications and authors. This web site is the primary source of data used by MGExplorer and the only alternative for accessing that particular dataset.

A. Methods and Metrics

We use the Keystroke-Level Model GOMS (KLM-GOMS) [28] analysis to simulate the effectiveness and efficiency of MGExplorer by estimating the average time (error-free) an expert user would take to solve the given task. To perform the analysis, we decompose the exploration flow in atomic actions such as cognitive (e.g., making a choice) and motor tasks (e.g., typing a word). We use the notation and temporal estimation of actions (for an average typist) proposed by Card et al. [29], in which: **K** refers to pressing a key or a button (ETA: 0.2s); **P** refers to pointing with the mouse to a target on the display (ETA: 1.1s); **M** refers to mentally preparing for a task (ETA: 1.35s); and **H** corresponds to homing the hand on the mouse (ETA: 0.4s).

B. Results

Tables III and IV present the atomic actions that compose each exploration scenario. Scenario 1 contains 27 atomic actions (8 clicks, 9 pointing, and 10 mental operations), which would take an expert user 25 seconds to complete. Scenario 2 contains 34 atomic actions (9 clicks, 12 pointing, and 13 mental operations), resulting in an estimated completion time of 32.55 seconds. This time difference could be explained by the adopted strategy to identify author A's most recurrent co-author in the GlyphMatrix, which required 9 extra pointing and clicking actions in a settings panel to sort the authors in a way that would ease the visual search within the matrix.

The HAL interface does not provide the users with counting authors' publications or co-authors, which means that they must calculate it by themselves. We estimate that the task would take at least 274 atomic actions (114 mental operations, 54 clicks, 54 pointing, and 52 homing actions). The interface allows filtering authors by name using a list of interactive alphabet letters. Upon selecting a letter, the system shows a list of authors whose names start with that letter. The user must then search for the author of interest within each list and click on it to explore their publications, co-authorship, and other related information. To identify the author with the most co-authors (author A), the user must visit every author's page and count their co-authors; these numbers must be compared to obtain the one with the most co-authors. The interface does not provide information on the actual number of authors under each alphabet letter, but we do know that there is at least one author for each letter, which means that the user has at least 26 authors to compare. Thus, for the sake of simplicity, we estimate the completion time of this task assuming that the user must perform 26 comparisons. Once author A has been identified, the user would revisit this author's page using the same filtering strategy, where they would compare the co-publications among their co-authors to identify the most recurrent one (author B) and the number of publications they have together (that information can be found next to co-authors' names). On author A's page, the user can also retrieve author A's total number of co-publications; this information is then compared to the number of co-publications with author B to determine the impact of this co-authorship on the total number of publications co-authored by author A. Due to the difficulty of identifying the author with the most co-authors, a process consisting of at least 26 loops taking around 10.1 seconds each, this scenario would take about 278.7 seconds (4.6 minutes) in total to be performed.

[4]https://hal.archives-ouvertes.fr/

TABLE III
Set of atomic actions performed in Scenario 1. Color encodes different visualization techniques: NodeEdge Diagram, IRIS, Bar chart, and Two Bar charts

Op	Description	Time
M	Identify the author with the most co-authors (**author A**)	1.35
P	Move the cursor to the node representing author **A**	1.1
K	Right-click on the node to further explore the data attributes	0.2
M	Choose a suitable information visualization technique	1.35
P	Move the cursor to the IRIS option	1.1
K	Click to select the IRIS visualization technique	0.2
M	Search for the longest bar in the IRIS, which refers to the most recurrent author (**author B**)	1.35
P	Move the cursor to the co-author's name	1.1
K	Click on the co-author's name to center it in the IRIS	0.2
K	Right-click on the bar to explore co-publications with author **B**	0.2
M	Choose a suitable visualization technique	1.35
P	Move the cursor to the Bar chart option	1.1
K	Select the Bar chart to explore the co-publications of both authors	0.2
M	Identify in the Bar chart the number of co-publications between the authors	1.35
M	Define a strategy to compare this result with the count of publications co-authored by author **A**	1.35
P	Move the cursor to the NodeEdge diagram	1.1
M	Search for the node representing author **A**	1.35
P	Move the cursor to the node representing author **A**	1.1
K	Right-click on the node to further explore the data attributes	0.2
M	Choose a suitable information visualization technique	1.35
P	Move the cursor to the Bar chart option	1.1
K	Click to select the Bar chart visualization technique	0.2
M	Determine how to compare the data in both Bar chart	0.2
P	Move the cursor to the title-bar of the Bar chart	0.2
K	Click on the title-bar of the Bar chart to select the panel	0.2
P	Move the Bar chart next to the Bar chart with co-publications of authors **A** and **B**	1.1
M	Compare the information in both Bar charts	1.35
Total exploration time in seconds (estimated)		**25**

TABLE IV
SET OF ATOMIC ACTIONS PERFORMED IN SCENARIO 2. COLOR ENCODES DIFFERENT VISUALIZATION TECHNIQUES: NODEEDGE DIAGRAM, GLYPHMATRIX, BAR CHART, AND BAR CHART AND GLYPH MATRIX

Op	Description	Time
M	Identify the author with the most co-authors (**author A**)	1.35
P	Move the cursor to the node representing author **A**	1.1
K	Right-click on the node to further explore the data attributes	0.2
M	Choose a suitable information visualization technique	1.35
P	Move the cursor to the GlyphMatrix option	1.1
K	Click to select the GlyphMatrix visualization technique	0.2
M	Define a strategy to compare the glyphs in the matrix's cells	1.35
P	Move the cursor to the T button in the GlyphMatrix's title-bar	1.1
K	Click on the button to open the settings panel	0.2
M	Determine which settings to apply	1.35
P	Move the cursor to the combo box under "Order by"	1.1
K	Click on the combo box to display the options	0.2
M	Identify the option the suitable option	1.35
P	Move the cursor to the option "Qt Publication"	1.1
K	Click to select the option "Qt Publications"	0.2
M	Identify how to hide the settings panel	1.35
P	Move the cursor onto the X button	1.1
K	Click on the button to hide the settings panel	0.2
M	Identify the most recurrent author (**author B**)	1.35
P	Move the cursor to the intersection cell of authors **A** and **B**	1.1
M	Identify the number of co-publications between authors **A** and **B**	1.35
M	Define a strategy to compare these results with the number of author **A**'s co-publications	1.35
M	Identify the matrix's row that correspond to author **A**	1.35
P	Move the cursor to the author's name	1.1
K	Right-click on the author's name	0.2
M	Choose a suitable visualization technique	1.35
P	Move the cursor to the Bar chart option	1.1
K	Select the Bar chart technique	0.2
M	Define a strategy to compare the Bar chart and the tooltip	1.35
P	Move the cursor onto the title-bar of the Bar chart	1.1
K	Click on the title-bar of the Bar chart to select the panel	0.2
P	Move the Bar chart next to the GlyphMatrix	1.1
P	Move the cursor to the intersection cell of authors **A** and **B**	1.1
M	Compare the information in the Bar chart and the tooltip	1.35
Total exploration time in seconds (estimated)		**32.55**

C. Discussion

The evaluation we performed allowed us to analyze the mental, physical, and temporal demands required from a user while using the MGExplorer tool. For the sake of simplicity, these demands are defined as follows: the *mental demand* refers to the amount of information one needs to have before the exploration and the information they must remember while exploring the data; the *physical demand* refers to the amount of physical interaction (e.g., pressing buttons or keys, moving the cursor from one point to another, etc.) necessary to accomplish the task; and the *temporal demand* refers to the time one has to dedicate into accomplishing the task.

Since the visualization interface displays every necessary piece of information either directly via colors, symbols, or text, or indirectly via interaction techniques, such as hovering over symbols to get data details on a tooltip, an expert user of the tool should be able to perform a particular analytical task with little mental effort.

While it is possible to explore co-authorship information using traditional listing interfaces, the user must compute a larger amount of information to obtain the same outcomes as using MGExplorer. Our temporal estimations showed that MGExplorer is very efficient to perform the selected task, requiring around 30 seconds of work, while one of the most straightforward exploration scenarios using the HAL interface is almost eight times slower. These high temporal and mental demands naturally increase the number of atomic actions and therefore the physical demand for performing the task in the HAL interface. Moreover, when using the visualization interface, the users can keep their hands on the mouse the whole time, reducing the need for homing actions, which are highly required when using the HAL interface, if we consider that the user would write down on a piece of paper (or type in a document) the counts of co-authors for further comparison.

When analyzing MGExplorer in the context of related works, we find Graph Trail [18] as the closest approach to ours because it allows the user to instantiate views through visual querying, although the interaction for doing that is different. It also represents the dependency between the views that display subsets of the dataset, thus providing provenance information. However, GraphTrail only supports querying on the original dataset to create new views and does not allow the user to hide or show the views created throughout the process, causing the visualization space to be quickly cluttered. This makes it difficult for the user to investigate alternative paths from a previous point in the exploratory history.

VI. CONCLUSIONS

In this paper, we presented MGExplorer, a visualization tool to assist the exploration of multidimensional networks using a chained views approach. Its effectiveness and efficiency were demonstrated through an assessment of co-authorship data exploration compared to a traditional interface. MGExplorer development is based on open-source libraries with a generic structure that provides a customizable layout by allowing the expansion of the visualization and interaction techniques palettes, benefiting both designers and users. Designers have a generic system that supports multiple visualizations while users can choose the ones suitable for their analysis.

The architecture of MGExplorer supports data from different sources and the exploration of data attributes of multidimensional networks from different application areas. It shows a data processing phase that detects the relationships between the dataset items and their attributes before visualization. Furthermore, a server supports the querying of multiple SPARQL endpoints, which allows the retrieval of data from different datasets for use in a view. Thus, the user can explore new hypotheses on the data and bring new data to the process to expand and improve the ongoing analysis.

The two exploration scenarios demonstrate that MGExplorer supports analytical tasks. Also, they demonstrate that an analytical goal can be achieved through different exploratory paths. Moreover, MGExplorer shows the analytical provenance in these exploratory paths. For our evaluation purposes, we did not consider the time for learning how to use the tools

(either MGExplorer or the website HAL). Moreover, we do not take into account variations of strategies that real users might develop to find information using *chained views*.

Although the exploration scenarios and our evaluation of the tool based on estimated completion time may be enough to support the claims about the effectiveness and efficiency of our approach compared to traditional interfaces, user-based evaluations are essential and should be performed to determine the usability and suitability of the MGExplorer interface. Future work includes developing user-based evaluations to investigate the usability of MGExplorer for exploring different multidimensional networks and its suitability for exploring co-authorship network data and assisting domain-related tasks.

The provenance information recorded by the MGExplorer is limited to the subsets of data and the visualizations used during the analysis. Thus, we intend to increase the variety of provenance information by recording, for example, atomic actions relevant to the reasoning and annotations made by users. Future work also includes the analysis of the resulting provenance data, e.g., to identify the most common usages of the system (standard choices of visualizations and instantiating order) according to different types of tasks, which could be used to introduce the system to new users, suggest some well-known analysis workflows, and to improve overall user experience. Furthermore, we could validate these usage patterns through user-based evaluations involving experts in different application domains. For example, the system could suggest different analysis workflows to perform particular tasks, and the users would evaluate whether and at which level that workflow responds to their needs and how to improve it.

VII. Acknowledgments

We would like to thank Valentin Ah-Kane and Olivia Osgart for their engineering work on the MGExplorer tool. We acknowledge the team members Alain Giboin, Fabien Gandon, and Catherine Faron-Zucker for all discussions that contributed for improving this research. We also acknowledge Brazilian funding agencies CNPq and CAPES (Finance Code 001) for financial support to R. Cava and C.M.D.S. Freitas in developing the first version of MGExplorer.

References

[1] B. Preim and K. Lawonn, "A survey of visual analytics for public health," *Computer Graphics Forum*, vol. 39, no. 1, pp. 543–580, 2020.

[2] S. Chen, L. Lin, and X. Yuan, "Social media visual analytics," *Computer Graphics Forum*, vol. 36, no. 3, pp. 563–587, 2017.

[3] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert, "A survey on visual analysis approaches for financial data," *Computer Graphics Forum*, vol. 35, no. 3, pp. 599–617, 2016.

[4] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[5] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*. IEEE, 2007, pp. 61–71.

[6] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for using multiple views in information visualization," in *Proc. Working Conference on Advanced visual interfaces*, 2000, pp. 110–119.

[7] J. and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 495–513, 2013.

[8] S. Liu, D. Maljovec, B. Wang, P. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249–1268, 2017.

[9] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "The state of the art in visualizing dynamic graphs," in *Eurographics conference on Visualization (EuroVis)–State of The Art Reports*, R. Borgo, R. Maciejewski, and I. Viola, Eds. Eurographics Association, 2014, pp. 83–103.

[10] C. Nobre, M. Streit, M. Meyer, and A. Lex, "The state of the art in visualizing multivariate networks," *Computer Graphics Forum (EuroVis)*, vol. 38, pp. 807–832, 2019.

[11] J. Leng, *Handbook of Research on Computational Science and Engineering: Theory and Practice: Theory and Practice*. IGI, 2011, vol. 2.

[12] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 31–40, 2015.

[13] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch, "Survey on the analysis of user interactions and visualization provenance," *Computer Graphics Forum*, vol. 39, no. 3, pp. 757–783, 2020.

[14] C. T. Silva, J. Freire, and S. P. Callahan, "Provenance for visualizations: Reproducibility and beyond," *Computing in Science & Engineering*, vol. 9, no. 5, pp. 82–89, 2007.

[15] T. Jankun-Kelly, K.-L. Ma, and M. Gertz, "A model and framework for visualization exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 2, pp. 357–369, 2007.

[16] C. Viau and M. J. McGuffin, "Connectedcharts: explicit visualization of relationships between data graphics," *Computer Graphics Forum*, vol. 31, no. 3pt4, pp. 1285–1294, 2012.

[17] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit, "Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2023–2032, 2014.

[18] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson, "Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, pp. 1663–1672.

[19] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim, "Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 120–130, 2017.

[20] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink, "Analytic provenance: process+ interaction+ insight," in *CHI Extended Abstracts Human Factors in Computing Systems*, 2011, pp. 33–36.

[21] Y. B. Shrinivasan and J. van Wijk, "Supporting exploration awareness in information visualization," *IEEE Computer Graphics and Applications*, vol. 29, no. 5, pp. 34–43, 2009.

[22] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "Vistrails: Visualization meets data management," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 745–747. [Online]. Available: https://doi.org/10.1145/1142473.1142574

[23] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit, "From visual exploration to storytelling and back again," *Comput. Graph. Forum*, vol. 35, no. 3, p. 491–500, Jun. 2016.

[24] F. Gandon, "A Survey of the First 20 Years of Research on Semantic Web and Linked Data," *Revue des Sciences et Technologies de l'Information*, Dec. 2018.

[25] R. Cava, C. M. D. S. Freitas, and M. Winckler, "Clustervis: visualizing nodes attributes in multivariate graphs," in *Proceedings of the Symposium on Applied Computing*, 2017, pp. 174–179.

[26] R. Cava, C. M. Freitas, E. Barboni, P. Palanque, and M. Winckler, "Inside-in search: an alternative for performing ancillary search tasks on the web," in *Latin American Web Congress*, 2014, pp. 91–99.

[27] R. Cava and C. D. S. Freitas, "Glyphs in matrix representation of graphs for displaying soccer games results," in *The 1st Workshop on Sports Data Visualization. IEEE*, vol. 13, 2013, p. 15.

[28] J. Sauro, "Measuring u," *Denver, Colorado*, 2010.

[29] S. K. Card, T. P. Moran, and A. Newell, "The keystroke-level model for user performance time with interactive systems," *Communications of the ACM*, vol. 23, no. 7, pp. 396–410, 1980.