This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

On Network Structural and Temporal Encodings: A Space and Time Odyssey

Velitchko Filipov *Member IEEE,* Alessio Arleo *Member IEEE,* Markus Bögl, Silvia Miksch *Member IEEE*

Abstract—The dynamic network visualization design space consists of two major dimensions: network structural and temporal representation. As more techniques are developed and published, a clear need for evaluation and experimental comparisons between them emerges. Most studies explore the temporal dimension and diverse interaction techniques supporting the participants, focusing on a single structural representation. Empirical evidence about performance and preference for different visualization approaches is scattered over different studies, experimental settings, and tasks. This paper aims to comprehensively investigate the dynamic network visualization design space in two evaluations. First, a controlled study assessing participants' response times, accuracy, and preferences for different combinations of network structural and temporal representations on typical dynamic network exploration tasks, with and without the support of standard interaction methods. Second, the best-performing combinations from the first study are enhanced based on participants' feedback and evaluated in a heuristic-based qualitative study with visualization experts on a real-world network. Our results highlight node-link with animation and playback controls as the best-performing combination and the most preferred based on ratings. Matrices achieve similar performance to node-link in the first study but have considerably lower scores in our second evaluation. Similarly, juxtaposition exhibits evident scalability issues in more realistic analysis contexts.

Index Terms—Human-centered computing–Visualization–Graph drawings, Empirical studies in visualization

1 Introduction

A ''graph" or "network" (the two terms will be used interchangeably in this paper) is a data structure representing a set of nodes and their relationships, represented by edges between them. Due to its flexibility and relevance to several application domains, the problem of how to visually represent such a structure has been attracting interest since the 1960s [1]. With the turn of the millennium, Herman et al. [2] first introduce a shift of perspective from "graph drawing" to "network visualization", mentioning how the latter deals with much larger graphs than before—a necessary adaptation to the rise in size of scientifically interesting networks. Around the same time, the increased availability of time-dependent datasets contributed to the increase of research interest in dynamic network visualization [3]. The paradigm shift from static to dynamic data represented a natural and necessary step forward to tackle the upcoming challenges presented by large and fast-evolving datasets, effectively anticipating and preparing for the advent of the "Big Data" era. Thanks to its constant evolution and growth for the last 20 years, dynamic network visualization is nowadays considered a mature and thriving field of research [3].

The visualization of a dynamic network can be considered, at its core, the problem of concurrently representing the graph topology, that is the underlying network structure, along with the temporal facet [4]. Kerracher et al. [5] introduce a design space for dynamic network visualization techniques along these two dimensions: the network *structural representation* (topology) and *temporal encoding* (dynamics). This two-dimensional design space is expressive enough to characterize the majority of existing dynamic network visualization approaches.

There is extensive literature on studies designed to evaluate different graph representations for typical exploration tasks on static networks (see, e.g., [6], [7]). Similar studies have been conducted for dynamic approaches, however they compare different temporal encodings for the same structural representation (node-link diagrams for the most part, see Section 2). This also comes as a consequence of the limited number of dynamic network visualization approaches that have matrices as their base graph representation [3] (see, e.g., [8], [9]). Furthermore, existing user studies in this context also incorporate simple interaction methods to support the network exploration (see, e.g., [10], [11]) Empirical evidence about the performance and preference for different dynamic network visualization approaches in our design space is still scattered over different studies, experimental settings, procedures, differnt interaction implementations, and tasks. This has two consequences. On the one hand, it makes it more difficult to assess and compare the different techniques in our design space, even on the same set of graph analysis tasks. On the other, it is not straightforward to identify which aspects of such techniques to select when building a network visualization system for the analysis of a dynamic real-world dataset.

Our Contribution. In this paper, we contribute a two-step evaluation to fill these gaps in the literature. First, we design, conduct, and discuss the results of a user study aimed at comparing different network structural representations, temporal encoding techniques, and interactions (Section 5). Second, we refined a set of visualization approaches based on the outcomes of the first study. We then evaluate these in a heuristic-based qualitative study with visualization experts to extract knowledge and obtain insights in a

• *All authors are with the Centre for Visual Analytics Science and Technology (CVAST), TU Wien, Vienna (AT).*
E-mail: {name.surname}@tuwien.ac.at

realistic analysis scenario (Section 6). Finally, we discuss the findings and takeaways across both evaluations (Section 7) and derive an overall conclusion (Section 8). We outline the following contributions resulting from our two-step evaluation:

- We assessed the accuracy, response times, and preference of node-link diagrams and adjacency matrices for dynamic graph tasks and concluded that the former performs better for high-level tasks (overview, estimation, higher-level structures), whereas the latter was more accurate for low-level tasks (identifying nodes, edges, and timeslices).
- Our study shows the influence that interaction techniques have on response times and accuracy regardless of the network's structural representation. Providing interactions significantly increases the response times, however, at the same time increasing the accuracy of the responses.
- The study results suggest that animation and animation with playback significantly outperformed the other techniques in our design space consistently.
- Our results show that node-link diagrams are the most preferred structural representation, as well as the one associated with the highest accuracy and lowest response times from both studies.

A preliminary version of this research has been presented at the International Symposium on Graph Drawing and Network Visualization (GD) 2022, selected by the Program Chairs and invited for publication in TVCG. This extended journal version contains revised writing and experiment description, a new qualitative evaluation, and a discussion of takeaways from both studies.

2 Related Work

We outline recent related studies conducted along the two dimensions of the design space introduced by Kerracher et al. [5].

Structural Representations. In graph drawing literature, several studies assess the readability, task performance, and effects of aesthetic criteria on human cognition of different graph structural encodings (e.g., [7], [11], [12], [13], [14], [15], [16], [17]). Okoe et al. [7], [11] conduct comparative evaluations between node-link and matrix representations on a large scale (~ 800 participants). Their results show that node-link diagrams better support memorability and connectivity tasks. Matrices have quicker and more accurate results for tasks that involve finding common neighbors and group tasks (i.e., involving clusters). Concurrently, Ren et al. [14] conduct a large-scale study (~ 600 participants) comparing the readability of node-link diagrams against two different sorting variants of matrix representations on small to medium social networks (~ 50 nodes). Their findings do not differ significantly from the ones by Okoe et al. [11], suggesting that node-link provided a better implicit understanding of the network, with lower response times and higher accuracy than matrices. However, the gap between the two tended to reduce as the size of the graph increased. Abdelaal et al. [18] conduct a crowd-sourced study (~ 150 participants) where bipartite layouts are compared with node-link and matrix-based representations on their performance on overview tasks for large graphs (~ 500 nodes) and detail task for smaller ones. Their findings suggest that matrices are the most reliable across all tasks, also providing evidence of the positive effect of bipartite networks in exposing the network structure.

Temporal Encodings. One of the most studied problems concerning dynamic network visualization, is the ability of participants to retain a "mental map" of the graph while investigating its evolution [10], [19], [20], [21], [22]. Archambault and Purchase investigate the effect of drawing stability on the node-link graph representation coupled with animation and small multiples [10], [21]. Drawing stability proved to have a positive effect on task performance, with animation able to improve over the timeline in low-stability scenarios. Ghani et al. [23] investigate the perception of different visual graph metrics on animated node-link diagrams. Results suggest that animation speed and target separation have the most impact on performance for event sequencing tasks. Linhares et al. [24] compare four different approaches for visualization of dynamic networks, namely the Massive Sequence View [25] (timeline-based), the Temporal Activity Map [26], and animated node-link and matrix diagrams. While all techniques reached satisfactory results, the animated node-link was the favorite choice of the participants. Even though matrix-based approaches are included in this study, it does not exhaustively cover all the possible combinations of our design space. Filipov et al. [27] conduct an exploratory study comparing different combinations of structural and temporal representations. The results suggest that tasks with matrices were completed quicker and more accurately, the participants preferred matrices with superimposition, and juxtaposition was among the least preferred approaches. However, these results comparing matrix-based approaches with node-link diagrams and their temporal encodings require further confirmation and formal statistical analysis. Overall, related literature shows that the perception of different temporal encodings has been mainly investigated on node-link diagrams, with few papers focusing on the other combinations of structural and temporal encodings. In this sense, our paper constitutes an effort in understanding whether the differences between node-link and matrix representations still hold in a dynamic scenario, what is the efficacy of the temporal representations, and how effective (and how important) is it to include interactions when designing such approaches.

3 Dynamic Graph Visualization

We refer to a dynamic graph Γ as a sequence of individual graphs each one representing its state at a specific point in time: $\Gamma = (G_1, G_2, ..., G_k)$; we denote the individual G_x as a dynamic graph *timeslice*. We now briefly describe the different structural and temporal encodings, detailing their implementation in our experiments.

3.1 Network Structural Representations

The structural dimension focuses on the challenges of laying out a graph to visually present the relationships between elements in an understandable, accurate, and usable manner [5].

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

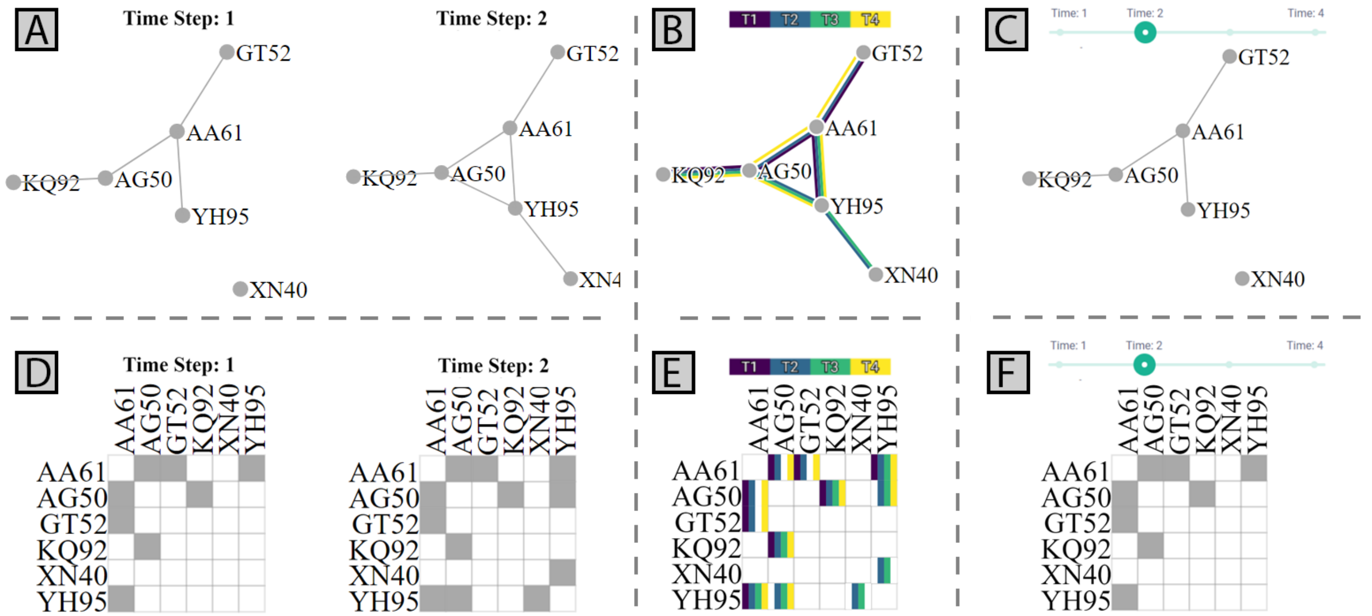

Fig. 1: Network structural and temporal encodings: Juxtaposition (A,D), Superimposition(B,E), and Animation with Playback Controls (C,F)

Node-Link (NL) diagrams present the relational structure of the graph using lines to connect the entities that are depicted using circles, whose coordinates on the plane are computed using specialized algorithms. In our study, we compute the NL layouts using the force-directed implementation of *d3js* [28] by aggregating all the available timeslices into one graph for which a drawing is computed. This process of *aggregation* [29] is simple to implement and provides a stable layout throughout the sequence of timeslices, at the expense of the quality of individual layouts. In this paper, we assume to have the complete time sequence available, therefore the drawing can be computed based on both *past* and *future* timeslices (an *offline* drawing approach [30]). This is opposed to an *online* scenario where the layout can be computed only based on *past* timeslices (e.g., when dealing with streaming data). We refer to the following for a broader discussion on dynamic network layout algorithms [3], [31], [32], [33], [34], [35], [36], [37], [38].

Adjacency Matrices (M) visualize the network as an $n \times n$ table. A non-zero value in the cell indicates the presence of an edge between the nodes identified by the corresponding row and column. In our study, we order the rows and columns lexicographical according to the node's label. This ordering would appear immediately familiar to the user without the need for further onboarding, and has been used in previous studies [15]. More advanced reordering methods exist [39], however, matrix reordering is still under-investigated in a dynamic context and we include this aspect in our secondary study design.

3.2 Network Temporal Encodings

In dynamic networks, the temporal dimension plays an important role in the analysis process and requires special attention to enable effective exploration and a better understanding of the behavior of the network [40].

Superimposition (SI) encodes the temporal dimension of the network in the same screen space by overlaying the timeslices (see, e.g., [33], [37]) or making use of explicit encoding (see., e.g., [41], [42]). In our study, we represent the temporal information in SI using colorblind-friendly color palettes [43]. In NL, we generate multiple parallel edges between the nodes, one for each timeslice where the edge is present, and color-code them individually. In M we subdivide each cell uniformly into rectangles, each representing the existence of that edge during that timeslice, which is colored similarly (see Figure 1 B-E).

Juxtaposition (JP) represents the graph's temporal dynamics as distinct layouts, each with dedicated screen space, similar to the small multiples approach by Tufte [44] (see Figure 1A,D). In our study, we generate one diagram per timeslice and arrange them adjacent to each other.

Animation with Playback Control (ANC) uses a time slider to control the state of the animation and move to any of the available timeslices in no particular order (see Figure 1-C,F). The transition always happens between the start and end time slice that is selected, intermediate ones are not considered. This enables a more fine and controlled exploration and analysis compared to animation, where speed and time progression are typically fixed.

Auto Animation (AN) depicts the change of the graph over time as smooth transitions between subsequent timeslices. Differently from ANC, with AN it is not possible to skip forward or navigate backward in time and it automatically goes over each of the timeslices in a sequence.

4 Evaluation Process

In order to understand the effect on performance and preference that different combinations of network structural and temporal encoding have on graph temporal tasks, we construct a two-step evaluation process.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

T#	Low-level	High-level
T1	At which time step is the relationship between {source} and {target} introduced for the first time?	At which time step does the clique between {nodes} appear for the first time?
T2	Sum up the changes (additions and removals) of {node}'s degree across all time steps.	Calculate the change of the clique's size between {nodes} across all time steps.
T3	At which time step does the node {node} have its highest degree?	Consider the set of nodes {nodes}. Find the size of the largest maximal clique across all the time steps between the given nodes.

TABLE 1: The test questions (trials), per task (rows) and entity type (columns).

In our first study (see Section 5), we opt for a controlled user study which would exhaust all possible combinations of our design space on the same predefined set of tasks. We also evaluate the impact of simple interactions by introducing their presence as a study condition. We perform a complete statistical analysis of the quantitative results of the study (i.e., response times and error rates), which we use to test our research hypotheses.

In the second study (see Section 6), we move beyond low-level benchmarks for predefined tasks and investigate how our proposed visualizations can be used to gain insights about a real-world dataset [45], also addressing some of the most important limitations that we identified in our first user study. We aim at exploring the capabilities of the techniques in the understanding of data and insight generation, by performing a heuristic evaluation meant to determine and compare the "value" of each visualization following the methodology proposed by Wall et al. [46]. We refine our initial selection of techniques by considering the results of the statistical tests as well as other qualitative information from the previous study (i.e., users' feedback and preferences) narrowing them down to two temporal encodings for both structural representations in our design space, for a total of four techniques (out of the initial 16). Moreover, we consider the feedback to enhance the interactions available for each of the techniques, simulating the use of more advanced visualization tools in the context of a real analysis scenario.

In the following sections, we present both studies in detail, discussing the study design, results, and limitations. We summarize the lessons learned in Section 7.

5 Study 1: User Evaluation

In this section we present the structure of our first study, describe the study design, including our tasks, research hypotheses, stimuli, and procedure.

Tasks. The tasks used in our experiment are available in Table 1. We picked one task for each category of temporal feature in the taxonomy proposed by Ahn et al. [47], namely, *Individual Temporal Features* (**T1**), *Rate of changes* (**T2**), and *Shape of changes* (**T3**). We selected the most common tasks referenced in the taxonomy and included in our experiment these tasks for both low- (nodes and edges) and higher-level (cliques) entities.

H #	Research Hypothesis
H1	M have lower response times and higher accuracy for all tasks compared to NL diagrams, regardless of the temporal encoding.
H2	From all temporal encoding techniques, SI has the lowest response times and highest accuracy, regardless of the structural representation.
H3	Providing interaction techniques increases the response times but not the accuracy.
H4	M have lower response times and higher accuracy for tasks on low-level entities and NL diagrams have lower response times and higher accuracy for tasks on higher-level entities, regardless of the temporal encoding.
H5	The combination M+SI results in the lowest response times and highest accuracy compared to other combinations of network structural and temporal encoding.

TABLE 2: The research hypotheses that were evaluated in our experiments.

Research Hypotheses. We base our research hypotheses on the proposed tasks and we report them in Table 2. The research hypotheses **H1**, **H2**, **H4**, and **H5** are derived from the observations and results of our previous exploratory study [27] (see also Section 2). While the focus of this experiment is centered around the *visual* encoding combinations within our design space, **H3** is intended to investigate the effects of common interaction techniques in this context. We conjecture that these increase response times over visual inspection alone without a significant impact on accuracy. This research hypothesis is determined empirically from previous work [27] where participants performed tasks on dynamic network visualizations within a similar design space as the one in this paper *without* the support of interaction methods. We assume that providing interactivity would have an impact on the response times, due to the time needed for the users to accept and then adopt it. Similarly, based on prior observations, we also assume that the benefit of providing interactions will not be associated with a significant increase in the accuracy of the tasks. In **H4** we conjecture that following the evolution of a cluster or clique is more difficult with M compared to NL. This assumption is derived from the results of our previous study where participants focused on low-level tasks (i.e., individual nodes and their temporal features [47]). Since in this study, the participants must track several elements at once, we expect this would be easier to achieve with NL as the nodes are drawn closer together, compared to M.

5.1 Interactions

The interactions we implement are meant to support network exploration. The following apply regardless of the temporal encoding: (i) zooming and panning (both for M and NL); (ii) hovering over a M cell highlights its corresponding row (from the left) and column (from the top); (iii) in NL, nodes can be moved by dragging in order to de-clutter some denser areas of the drawing. Moreover, for AN only and regardless of the structural representation, the time between consecutive timeslices can be increased (7 sec maximum) or decreased (1 sec minimum). This selection should not favor any specific combination of structural and temporal encoding techniques over the others. Zooming,

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

panning, and node rearrangement are commonly available in graph exploration software, like *Gephi* [48]. M mouse-over was also used by Okoe et al. [11]. AN speed could also be manipulated in the study by Archambault and Purchase [20].

5.2 Experiment Setting

Stimuli. We generated 24 different scale-free random [49] graphs ($35 \leq |V| \leq 45, 46 \leq |E| \leq 71$) with the *NetworkX* python library [50], [51]. We chose this category of networks as they resemble real-world data examples of scientific interest (e.g., the world-wide-web, authors' co-citation networks [52]). We augmented each graph with 4 timeslices as follows. A single timeslice was created from the input graph by randomly deleting edges; the process is repeated for each of the 4 required. Then, they are arranged to form a sequence (in no particular order), so that we simulate the temporal dynamics of edge addition/removal. Finally, we split the datasets into two different types: 12 graphs with cliques and 12 without. As the input graphs did not naturally include cliques, they were introduced artificially when necessary. 5 random nodes in each graphs were randomly selected, and then new edges were added to form the clique in one or more random timeslices (simulating the clique forming and breaking). The size of the graphs is comparable with the majority of empirical studies on graph visualization [14], [53].

Trials. Each of the tasks is applied to all combinations of structural and temporal encodings of interest in our study (see Section 3) resulting in 48 unique trials: $3(task\ types) \times 2(entity\ types) \times 2(network\ encodings) \times 4(temporal\ encodings)$. The entity types refer to either low-level (node or link) or high-level (cliques or clusters) components of the network [47]. The order of the trials during the study is randomized in order to mitigate learning effects. The participants take part in the online experiment by completing the trials prepared using SurveyJS [54].

Study Design. Our experiment follows a between-subject arrangement: all participants complete the same entire set of 48 trials on the same graphs, but are exposed to one of two conditions, either *without* (Group *A*) or *with* (Group *B*) the support of the interactions discussed in Section 5.1. Participants were randomly assigned to the two groups, with the majority (75%) in Group *B*. This subdivision is justified by the fact that only one of our research hypotheses (**H3**) requires participants *not* to take advantage of interactions. Therefore, we design the experiment to have a higher number of participants with interaction support. We estimated a split of 25% over the expected number of participants, as a sufficient size for Group **A** to obtain statistically significant results, see Sections 5.3 and 5.4 for further details. For each trial, we ask the participant to provide a confidence score of their answer using a 5-point Likert scale (1 least confident - 5 most confident). At the end of the experiment, the participants express their thoughts in a text field (i.e., "Please enter any personal comments") about the encoding combinations they encountered and rank them on a 5-point Likert scale (1 least preferred - 5 most preferred).

Participants. For our study, we enrolled students who were part of a graduate course on information visualization design. To ensure that participants had a sufficient level of knowledge on the topic, we gave an online introductory lecture about the visualizations and the experiment modalities. Participation was optional and its performance did not impact the final grade of the students. The online setting was necessary to guarantee a sufficient number of participants while ensuring a safe social-distancing protocol. However, this also meant giving up control of the experiment environment (i.e., no control over the participants' setup, devices used, and resolution).

H#	Groups	MWU	T-Test	Binomial
H1	*(NL T1) vs (M T1)*	$\mathbf{0.0104}^{*b}$	$\mathbf{<0.001}^{***b}$	$\mathbf{0.0013}^{*b}$
	(NL T2) vs (M T2)	0.1579	$\mathbf{<0.001}^{***b}$	0.9313
	(NL T3) vs (M T3)	$\mathbf{<0.001}^{***b}$	$\mathbf{<0.001}^{***b}$	$\mathbf{0.0022}^{**b}$
H2	*(SI) vs (JP)*	$\mathbf{<0.001}^{***b}$	0.1065	0.166
	(SI) vs (ANC)	0.8662	0.1429	0.0883
	(SI) vs (AN)	0.2766	0.7751	$\mathbf{<0.001}^{***b}$
H3	*(Grp A) vs (Grp B)*	$\mathbf{<0.001}^{***}$	$\mathbf{<0.001}^{***}$	$\mathbf{<0.001}^{***}$
H4	*(M Low) vs (NL Low)*	$\mathbf{<0.001}^{***b}$	0.1392	$\mathbf{<0.001}^{***b}$
	(M High) vs (NL High)	$\mathbf{<0.001}^{***b}$	$\mathbf{<0.001}^{***b}$	0.4321
H5	*(M+SI) vs (M+JP)*	$\mathbf{0.0056}^{**}$	0.2567	0.2424
	(M+SI) vs (M+ANC)	0.6301	0.2989	0.0261
	(M+SI) vs (M+AN)	0.2766	0.6328	0.0646
	(M+SI) vs (NL+SI)	$\mathbf{0.0038}^{**b}$	$\mathbf{<0.001}^{***b}$	0.449
	(M+SI) vs (NL+JP)	$\mathbf{<0.001}^{***b}$	$\mathbf{<0.001}^{***b}$	0.1389
	(M+SI) vs (NL+ANC)	0.0088	$\mathbf{<0.001}^{***b}$	$\mathbf{<0.001}^{***b}$
	(M+SI) vs (NL+AN)	0.0331	$\mathbf{<0.001}^{***b}$	$\mathbf{<0.001}^{***b}$

TABLE 3: The results of the statistical test (p-values) for each research hypothesis. We mark the cells with * if $p < 0.05$, ** if $p < 0.01$, *** if $p < 0.001$. If multiple comparisons are performed, ^{b} indicates the Bonferroni correction [55].

5.3 Analysis Approach

We received a total of 76 submissions from as many participants, of which we removed 8 that either recorded anomalous response times (way too quick or long) or incorrect answers to control questions, suggesting participants trying to "game" the study. This resulted in a final set of 68 valid submissions that were used as the basis of our analysis. We provide further details as supplemental material.

For each question of our study, we collected the participants' answers, their corresponding response times, and confidence values. We ignored the group subdivision (Group *A* and *B*) for research hypotheses that did not focus on the presence of interactions in the visualizations (all except **H3**, see Section 5), as ANOVA tables do not show a statistically significant interaction effect between the independent variables for **H1**, **H2**, **H4**, **H5** (for more information we refer to [56]).

We conduct our analysis as follows, supported by Python libraries for statistical analysis [57], [58], [59]. We consider the structural and temporal encoding, the task type, entity type, and the groups (Group *A* and *B*) as *independent variables*, the response times and accuracy are taken as *dependent variables*. As the group subdivision is not even (25-75), we choose methods that are robust against these unbalanced designs [60], [61], [62], [63]. For each of the research hypotheses, we decomposed them into simpler hypotheses and executed multiple null hypotheses tests (see groups in Table 3) to find evidence for or against our research hypotheses. For multiple group comparisons, we

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

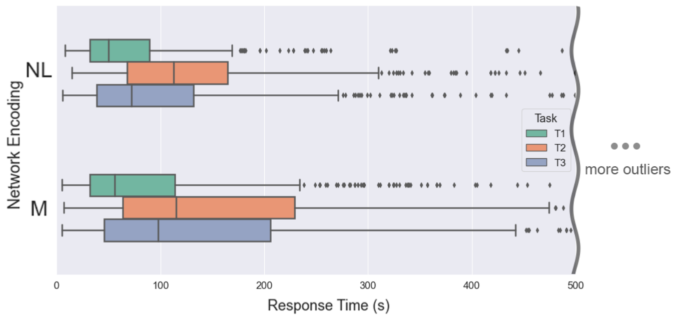

Fig. 2: **H1:** Box plot of response times for NL and M per task.

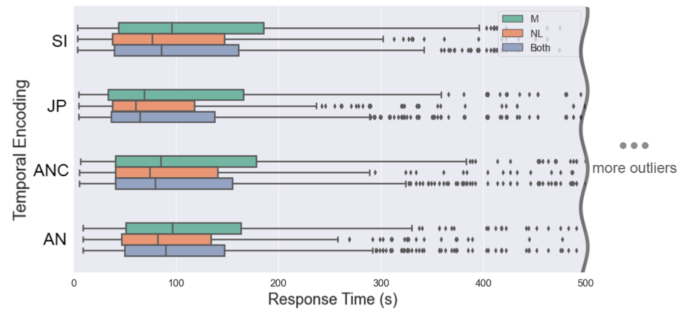

Fig. 3: **H2:** Box plot of response times for temporal and network representations.

took countermeasures by using the Bonferroni corrected alpha significance levels [55]. For the analysis, we consider and visually inspect response times and accuracy (number of correct answers ÷ total number of answers). To remove outliers from the data before the analysis we employ the inter-quantile range (IQR) [64]. We set the IQR lower ($q_1 - 1.5 \cdot \text{IQR}$) and upper ($q_2 + 1.5 \cdot \text{IQR}$) bounds at $q_1 = 0.25$ and $q_2 = 0.75$ as the outlier cut-off boundaries. This resulted in 116 trials (or 3.43%) being detected as outliers and omitted from the analysis.

The task response times in our experiment are not normally distributed. To mitigate this, we perform a Box-Cox transformation [65]. Visual inspection of the quantile-quantile (Q-Q) plots confirmed a normal distribution of the transformed data. This allows us to run parametric tests, specifically, ANOVA (see [56] for further information about the ANOVA tables) and T-tests [60], [61], [62], [63]. The standard ANOVA and T-tests are robust against such skewed distributions [66], [67], [68], therefore, we rely on them for our analysis as they both have more statistical power than non-parametric tests and detect significant effects if they truly exist. In the presence of a statistically significant difference (p-value < 0.05), we check, with T- and Mann-Whitney-U (MWU) tests, whether the significance held and visually explored the corresponding box plots to come to a conclusion. To evaluate our research hypotheses on accuracy, we also perform Binomial tests to detect statistical significance between the distributions.

5.4 Quantitative Results

H1. We presume, based on previous work [27], that M would perform better overall compared to NL for all tasks. Figure 2 depicts differences in response times between M and NL diagrams per task type. The results (see Table 3) indicate that NL is generally faster and more accurate than M. However, when looking at their differences per task we discover for **T1** that NL is significantly faster than M (NL: 73.49s, M: 97.93s), whereas M proves to be more accurate (NL: 74.9%, M: 80.7%). For **T2** the T-Test detects a significant difference in response times between NL and M (NL: 133.41s, M: 194.20s), however, in terms of accuracy they both perform similarly (NL: 52.5%, M: 52.7%). For **T3** NL representations significantly outperform M in terms of response times (NL: 107.32s, M: 175.92s) as well as accuracy (NL: 65.7%, M: 59.4%). Summarizing, the results suggest NL to generally have the lowest response times and higher accuracy compared to M for the proposed tasks. Thus, our results do not support H1.

H2. We assume SI to have the lowest response times and highest accuracy out of all the temporal encoding techniques. In our analysis, however, we do not detect any statistical significance in the comparisons shown in Table 3, with the only exception being JP, which has considerably lower response times than SI (see Figure 3). Concerning response times, JP has the lowest (118.32s), followed by AN (127.76s), SI (129.69s), and ANC (141.35s). We also run a paired T-Test comparing the temporal encoding approaches to check for statistical significance between pairs out of our initial research hypothesis and detect a significant difference between JP and ANC. In terms of accuracy, we discover a significant difference between SI (62.1%) and AN (68.6%). Whereas, between SI and JP (64.45%) or ANC (59.13%) there is no significant difference. We conjecture these results to be due to the graph's size and limited number of structural changes over time, which might favor AN as it is possible for participants to follow all changes during animation. Our analysis shows no evidence to support H2.

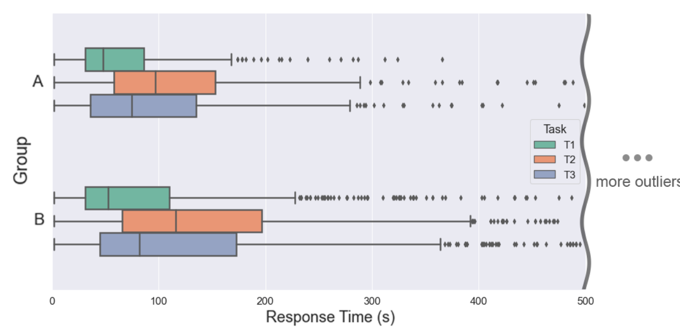

Fig. 4: **H3:** Box plot of response times for interaction groups per task.

H3. We conjecture that providing interactions influences the response times but not the accuracy. Our tests detect a significant difference (see Table 3) in the response times between group A (no interactions; 114.76s) and B (interactions; 163.83s). As we initially assume, the group with interactions is much slower in completing tasks than the group with no interactions (see Figure 4), however, the difference in accuracy is unexpected. The group with interactions is significantly more accurate than the one without (group A: 58%, group B: 65%). This suggests that interactions in-

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

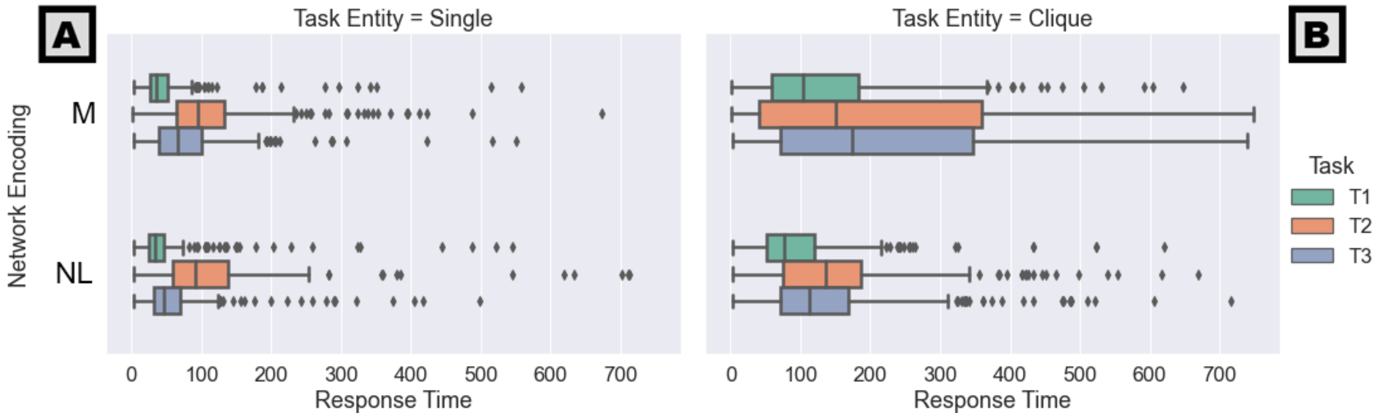

Fig. 5: **H4:** Box plot of response times for (A) single entities and (B) cliques.

deed increase response times, but at the same time provide the participants with a much better understanding of the visualized graphs and corresponding network dynamics regardless of the temporal encoding, therefore, leading to more accurate responses. The analysis shows that our results support H3 in terms of response times, but not accuracy.

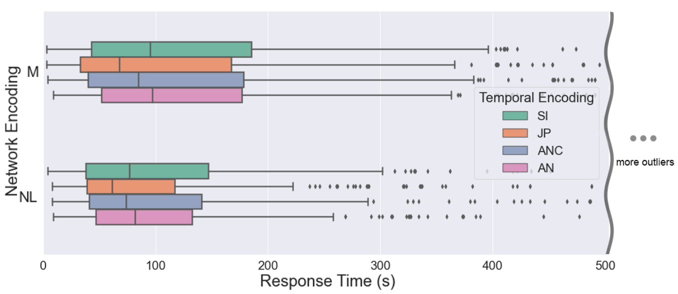

Fig. 6: **H5:** Box plot of response times for temporal and network representations.

H4. We formulate this research hypothesis to evaluate whether the response times and accuracy of M and NL representations are affected by the type of target entity in a dynamic context (*low-level* - individual nodes and edges; or *higher-level* - cliques), regardless of the temporal representation. For low-level entities, we do not detect any significant differences in the response times between network representations (see Table 3), both NL and M diagrams perform similarly. The results (see Figure 5) for tasks on low-level entities indicate that M has lower response times (NL: 97.08s, M: 90.24s), whereas for higher-level entities NL has significantly lower response times (NL: 146.66s, M: 245.2s). However, in terms of accuracy M is significantly better than NL for lower-level entities (NL: 82.1%, M: 86.4%). For the higher-level entities, NL and M representations perform quite similarly in terms of accuracy (NL: 42.1%, M: 41.3%) Based on these findings, the results suggest that H4 is partially supported.

H5. Finally, we want to assess the response times and accuracy for all possible combinations of network structural and temporal encodings. Our assumption is that M representations with SI temporal encoding have the lowest response times and highest accuracy. We compare M+SI to all other combinations of network structural and temporal encodings (see Figure 6). The results of the statistical tests yield significant differences in response times when comparing M+SI (154.53s) with M+JP (140.13s), NL+SI (105.25s), NL+JP (99.54s), NL+AN (108.8s), and NL+ANC (110.97s). Between M+SI (154.53s) and M+ANC (168.87s) and M+AN (160.62s) there is no significant difference in response times (see Table 3). In terms of accuracy, we detect statistically significant differences between M+SI (61.1%) and NL+ANC (51.8%) and NL+AN (71.4%). Whereas, the other combinations do not differ enough to warrant significance: M+JP (64%), M+ANC (66.4%), M+AN (65.5%), NL+JP (64.6%), and NL+SI (62.9%). From these results, the most balanced combination in terms of response times and accuracy is NL+AN followed by NL+JP. Therefore, we find no evidence supporting H5.

5.5 User Ratings and Feedback

We collect the participants' ratings per combination of network structural and temporal encoding along with textual feedback pertaining to their preferences and experience during the experiment (see Figure 7). There are no major differences in the preferences between the SI and JP encodings; ANC is the most preferred temporal encoding when coupled with an NL base representation. The NL representation is generally the most preferred approach, regardless of the temporal encoding. In terms of the participants' confidence, we observe that most participants seemed to be fairly confident in their answers across all approaches (see Figure 8). Most notably, the participants were most confident with NL+JP, followed by M+ANC, NL+ANC, and M+JP. There is general consensus that NL+SI was a very cluttered combination, whereas for M it performed a lot better and was easier to understand ("*SI was really confusing for some of the NL tasks but really useful for many of the M tasks*"). This is presumably due to the clutter generated by parallel edges crossings that occur in NL diagrams, which does not affect M. As in previous studies [27], the feedback on JP outlines that it requires participants to split their attention between multiple views in order to compare the temporal information. The ANC approach was preferred by the study participants for its flexibility due to the additional controls (i.e., time slider). AN was not considered to be a very good temporal encoding technique with the feedback being consistent across structural representations. Some participants

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

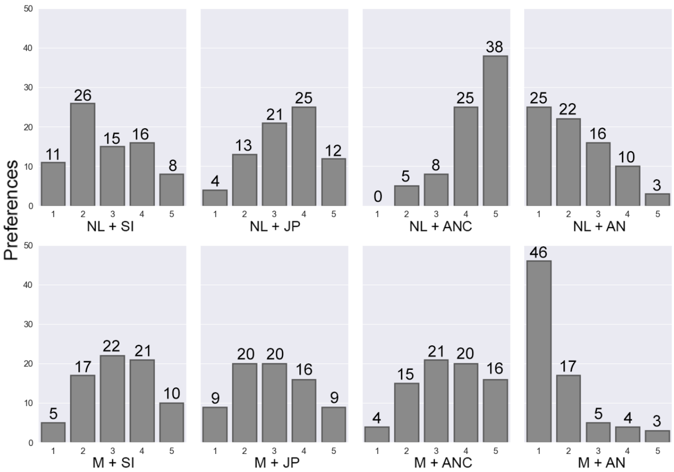

Fig. 7: Preferences per network and temporal encoding on a Likert scale (1-5).

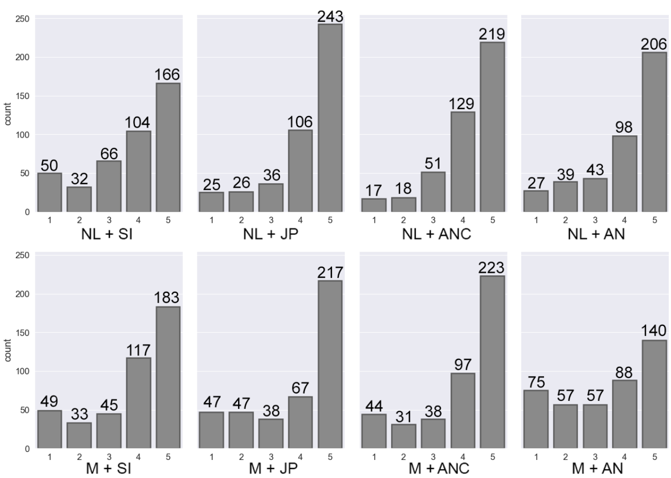

Fig. 8: Confidence per network and temporal encoding on a Likert scale (1-5).

commented that they needed to "*screenshot every timestamp to look at the different connections between the nodes*" and wait to watch the whole animation from the beginning. NL+AN, therefore, appears to be the least practical of the approaches, however, it also provides the best results. We conjecture this to be due to the size of the graphs and the number of structural changes occurring. M+AN is the lowest rated by the participants. The general consensus for AN is that it was difficult to keep track of the changes occurring between the nodes, requiring the viewer to memorize node positions and labels incurring a high cognitive effort to complete the tasks. Despite the aforementioned drawbacks, AN scales better to a larger amount of timeslices compared to SI and JP. Finally, the group with interactions had a better experience overall compared to the group without. The majority of the members of this group explicitly requested interactions to be implemented, supporting our findings concerning H3.

5.6 Limitations

In this experiment, the **size** of the graph was not considered when preparing the stimuli. Small graphs were chosen, both in the number of nodes/links and timeslices. M scales better to larger graphs than NL, while AN and ANC support a greater number of timeslices compared to SI and JP. Second, we chose simple, custom implementations for our structural and temporal encodings, disregarding more advanced solutions in literature (see Section 3). While this was done with the intention of testing the fundamental principles of the techniques in our design space, evaluating more sophisticated approaches might have significantly impacted the results. Finally, we focus on a selection of tasks from a taxonomy on network evolution analysis [47], other graph-based taxonomies could present relevant benchmarks for the proposed techniques.

In contrast to our previous exploratory study [27], the analysis of the results in this paper show that M-based approaches do not perform well in terms of response times and accuracy compared to NL diagrams. Our results also confirm the outcomes of similar studies evaluating the differences between structural representations in a static environment [7], [11]. Many of our research hypotheses are not supported and this challenges our opinions about the usefulness of M-based representations in a dynamic context. Therefore, building on the outcomes of this study, we conduct a further evaluation with the goal of investigating how well the techniques support extracting insights and gaining knowledge about the network's dynamics. We describe the experiment setting and results in our follow-up study (see Section 6), where we also aim at overcoming the limitations we identified.

6 Study 2: Heuristic Evaluation

In this section, we describe the structure of our second study, discuss the design, present our dataset, stimuli, and the evaluation procedure we follow.

Complementary to the previous study, we evaluate the techniques based on their potential to gain insights about the depicted dataset. For this purpose, we follow the heuristic evaluation methodology ICE-T by Wall et al. [46]. The goal of this evaluation is to determine the value of the best-performing visualizations from the first study: the value of visualization is defined as its capability of responding to data-driven questions, generating insights, and inspiring confidence in the potential results of the analysis [46].

6.1 Tested Techniques and New Interactions

We selected four combinations from our design space for this evaluation as the overall best performing from our first study (see Section 5). These are: NL+JP, NL+ANC, M+JP, M+ANC. Based on the user feedback, we modified the visualizations from the first study.

In terms of visualization, we improved the highlighting for both M and NL (see Figure 9). In NL diagrams hovering with the mouse over a node highlights it as well as its adjacent nodes. With M we highlight the entire column and row of the currently selected cell, providing the participants with a better view of intersecting cells. We depict the inactive nodes (i.e., not present at the current timeslice) in gray. Furthermore, we added tooltips for both structural representations showing the viewer the label(s) of the currently selected node or cell.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

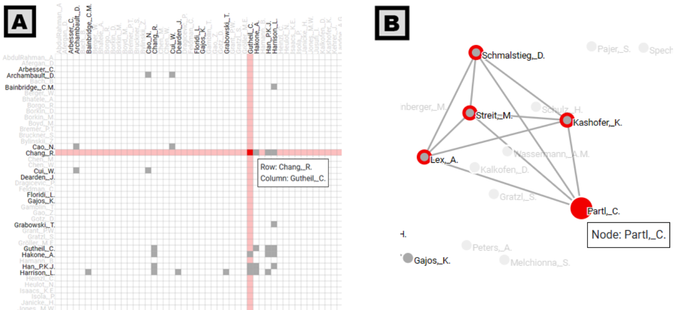

Fig. 9: Highlighting in the visualizations: (A) both the columns and rows of the selected cell in M-based representations are highlighted; (B) the selected node and its adjacent neighbors are highlighted for NL diagrams.

In terms of improved interaction techniques, for NL diagrams we implemented sticky nodes as a way for the participants to reorganize the layout of the network mitigating edge crossings in denser areas: each time a node is dragged around, its position becomes fixed and is not changed by the layout algorithm. For M-based representations, we included reordering algorithms, that are part of the ReorderJS library [69] (Leaf Order, Reverse Cuthill-McKee, Spectral, and Barycentric), with the aim of outlining structural patterns. Users can switch between the different reordering techniques at any point in the exploration. We also improved interaction with the temporal encodings as follows. For ANC we updated the animation to include both auto-animation as well as a time slider providing more interactive control. For JP we updated the layout to accommodate 8 timeslices so that all of these are visible on a single screen.

6.2 Experiment Setting

Stimuli. For this evaluation we used a real-world dataset describing co-authorships between authors in the Information Visualization community from 2008 to 2016, with 8 timeslices (compared to 4 in the first study) each spanning 1 year, obtained from a co-citation network [70]. We aimed at obtaining a graph about twice the size as the ones from the first study. We filtered the original input graph as follows: first, we stacked and ordered the graph nodes by descending degree; second, we popped the stack and included that node and its neighborhood in the filtered graph. This process continued until we reached a predefined threshold of 80 nodes, which we increased to 107 to avoid breaking any existing clique. This resulted in a total of 469 individual edge occurrences (avg. 58,6 per timeslice). In order to evaluate the capabilities of the visualizations to highlight invalid, unusual, or unexpected data cases [46], we inserted an additional 194 random edge occurrences, that brought up the average to 82,8 edges per timeslice. This is well within the typical scale of the graphs used in other evaluations [53].

Trials. In this study, we apply the heuristic evaluation by Wall et al. [46] aimed at evaluating the value of a visualization. This evaluation protocol entails an open-ended exploration of the data, where participants identify their own data-driven questions and find the corresponding answers. To initiate the analysis, we encouraged the participants to perform free-form exploration and browsing of the network and the timeslices, pointing out interesting insights they found (e.g., highly connected nodes, changes in relationships, reoccurrences, cliques or clusters that are formed, more interesting timeslices).

Study Design. We conducted the experiment as individual expert interviews remotely using a video conferencing platform (Zoom [71]) that lasted 60 minutes on average, preceded by a 10-minute introduction to the scope of our evaluation. Our visualizations were implemented using Angular [72] and d3.js [73] and hosted on a web server accessible to our participants from their own devices. The participants shared their screen content, which allowed us to record their activities and interactions as well as audio recordings. We kept a protocol of notes for each interview as well as reviewed the recordings after the evaluation sessions. Each participant spent about 15 minutes per visualization technique, exploring and interacting with the data, gaining insights, and voicing their thoughts about what they are doing, searching for, finding, or expecting to see. In the end, we asked them about their opinions on each visualization technique and to fill out one ICE-T survey [46] for each technique at their own convenience. We provided the surveys as online forms, where we would collect the results and calculate the overall score for each visualization technique afterward. Additionally, we added a field where the participants could explicitly provide any textual feedback pertaining to the visualization or interaction techniques, what they found useful, and what could be further improved (i.e., "Feel free to add any comments and feedback here."). Interviews were done in an uncontrolled setting. Participants used their own devices to complete the study by accessing the experiment online.

Participants. We recruited five participants, which is considered an appropriate number for an ICE-T survey stated by Wall et al. [46]. All participants are experts and have experience in both visual analytics and network visualization with a prominent publication track record in these fields. To ensure that they were all informed about the different modalities of our study we introduced the visualization and interaction techniques, dataset, and provided a brief explanation of patterns that might occur in M-based representations prior to the evaluation.

6.3 Analysis Approach

The ICE-T evaluation methodology used in this study [46] is structured hierarchically into 4 aspects relevant to visualizations (components), which are Insight, Time, Essence, and Confidence. Each component contains two to three visualization guidelines (intermediate level), with each encapsulating one to three heuristics. Heuristics represent how the visualization guidelines can be achieved (e.g., "*The visualization provides a big picture perspective of the data*"). These 21 heuristics are formulated as rateable statements asked directly to the participants (e.g., "*The visualization presents the data by providing a meaningful visual schema*"). Each participant provided a response to each one of the 21 heuristics in the questionnaire.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

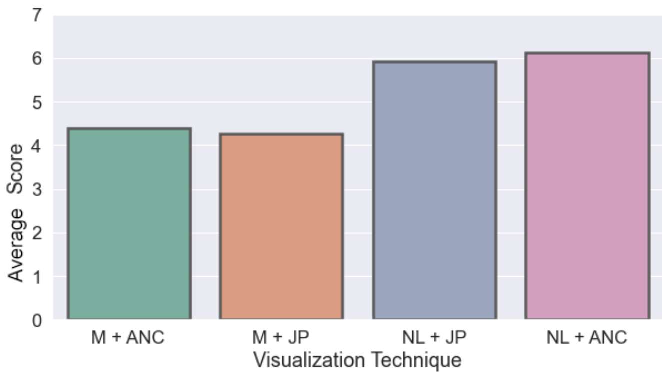

Fig. 10: Results of the ICE-T heuristic evaluation. This figure shows the average scores of each of the combinations of structural and temporal encodings evaluated in the study. The scores vary from 1 (Strongly Disagree) to 7 (Strongly Agree).

Heuristics are rated on a 7-point Likert scale with 1 being "Strongly Disagree" and 7 being "Strongly Agree". Additionally, there is a "Not Applicable" answer in case any of the participants thought the question does not apply (and excludes the answer from the score calculation). Scores are then averaged per component, whose average value is the visualization's overall score (or value). According to this methodology [46], a valuable visualization should obtain an overall average score of at least 5. To make the reported scores more transparent, we included the standard deviation of the ratings in our results together with the average scores (see Table 4). We include in the supplemental material the complete questionnaire and a detailed breakdown (including the individual responses) of the results we obtained. For more details about the ICE-T methodology, we refer to Wall et al. [46].

We report the quantitative results of this evaluation in Section 6.4. In addition, to the heuristic evaluation, we collected the participants' qualitative feedback and discuss it in Section 6.5.

6.4 Quantitative Results

The results of the evaluation (see Figure 10) indicate a clear preference of the participants for NL representations compared to M. This result is also consistent and confirms the findings of our first study, especially the ones concerning participants' preferences (see Sections 5.4 and 5.5). NL+ANC obtained a score of 6.15, whereas NL+JP was 5.96, while M did not get past 4.26.

Unsurprisingly, study participants found NL visualizations to provide a more natural and intuitive representation of the topology of the network, being more capable of highlighting structural patterns thus supporting insight generation and knowledge extraction. In turn, when using M representations a viewer must first decode the visualization in order to make sense of the relational data and structure, which takes more time and cognitive effort compared to NL.

When comparing the temporal encoding, a reoccurring trend from the first study (see Section 5), is that ANC is consistently preferred to JP. We presume this to be due to the nature of JP, which requires the viewer to continuously switch attention between the individual timeslices in order to trace an individual node, edge, and/or cell, and to observe its behavior over time. However, differently from JP, with ANC is not possible to compare distant (non-adjacent) timeslices.

We additionally inspected the scores of each visualization technique on a per-component basis (see Figure 11). The scores are mostly consistent throughout the four components, following the same trend as in Figure 10) with the exception of M+JP and M+ANC for the "Essence" (i.e., "*Live view or summarization of the dataset*" [46]), where M+JP scores better (4.36) than M+ANC (3.93). This may also come as a consequence of the divided opinion on M-based representations in a dynamic context, which is indicated by the high variance of both ANC and JP (see Figure 11). While the participants were interacting with the M-based representations, overall they favored ANC over JP, however, their opinions were somewhat divided if they found M useful for such analysis tasks or not.

The NL-based network visualization results indicate that across all components the ANC temporal encoding is consistently preferred to JP (see Figure 11). Furthermore, the participants' opinions and scores on NL diagrams are more consistent and in agreement compared to M representations as seen in Figure 11 (low variance for all NL techniques).

Overall, the best-performing combination of techniques (as well as the highest valued one), indicated by these results is NL+ANC (6.15), followed by NL+JP (5.96), M+ANC (4.46), and, finally, M+JP (4.26) (see Figure 10).

6.5 Qualitative Results

According to the study participants, with ANC it was easier to track changes occurring to a specific part of the graph (local changes), whereas for JP it was easier to observe more global changes happening over all the available timeslices and the entire graph regardless of the structural representation. With the support of our linked interaction techniques in JP, it was easier to compare distant timeslices compared to ANC, which most participants expressed being only useful to compare up to two neighboring timeslices. Thus, we believe the choice of temporal encoding, in this case, depends heavily on the granularity of the network analysis that participants are interested in (i.e., local vs. global behavior, distant vs. neighboring timeslices). Generally, the participants observed how ANC offered a lot more screenspace compared to JP for each timeslice, which is common for time-to-time vs. time-to-space visualization techniques [74].

Our decision of favoring layout stability to preserve the viewers' mental map both in NL and M was generally received favorably by the study participants. However, they also remarked that a balance between the layout stability and quality on a per-time slice basis may improve graph readability and insight generation - despite the cognitive impact on the users' mental map. In support of this claim, several participants argued that the existing reordering algorithms available for M representations were not able to emphasize topological patterns existing on individual timeslices as reordering was applied on the matrix representing the aggregated graph. Due to the nature of the

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

Technique	Component	Insight		Time		Essence		Confidence		Total	
	Score	AVG	STD	AVG	STD	AVG	STD	AVG	STD	AVG	STD
	M+ANC	4.65	1.22	4.25	1.84	3.93	1.44	4.73	1.73	4.46	1.53
	M+JP	4.33	1.46	3.85	2.06	4.36	1.91	4.53	1.96	4.26	1.78
	NL+ANC	6.22	0.82	6.00	0.84	6.14	0.83	6.13	1.02	6.15	0.87
	NL+JP	6.10	0.80	5.65	1.19	5.93	0.96	6.00	0.97	5.96	0.97

TABLE 4: Results of the ICE-T heuristic evaluation. The rows are the individual combinations of structural and temporal graph encodings that were evaluated. The columns are grouped per component and within each we calculate the mean and standard deviation of the participants' ratings. The right-most column shows the total for each technique.

publication dataset, where local neighborhoods can change drastically across adjacent timeslices, it was a challenging task to observe the global behavior of the network in M-based representations, regardless of their temporal encoding. This is also indicated in the results of M-based representations in the "Total" column of Table 4, where the standard deviation of the participants' ratings are significantly higher than those for the NL representations. Based on these results and our observations during the evaluation sessions this illustrates the diverging opinions and preferences that the participants had about M-based representations. Based on the interview videos, we also observed that all participants managed to identify smaller-scale structures and important or central nodes with the aid of matrix reordering algorithms. Using the M+ANC combination was more useful for detecting local changes compared to JP, which is consistent with the participants' general feedback about the temporal encodings. Based on the comments of the participants it became clear to us that M-based representations possess a steep learning curve in order to be decoded properly. However, when asked about the usefulness of M-based representations, there was also agreement there is potential to improve these by including more sophisticated interaction techniques, visual encodings, and reordering algorithms ("*M seems a bit nerfed*").

For NL+JP all participants expressed that the linked interactions and adjacency highlighting were extremely useful in order to observe changes occurring to a node and its neighbors over time. From the evaluation sessions we observed that NL+JP was regarded by most participants as providing a better view of the network and being effective at supporting cluster/clique identification, as well as, central (or bridge) nodes. However, there was consensus among participants on NL+ANC being the most intuitive approach contrary to their initial expectations. Prior to experimenting with the approaches, the participants assumed that NL+JP would be the most efficient technique for extracting insights. NL+ANC made it easier to identify persistent nodes and how clusters/cliques appear, disappear, or evolve over time due to the natural time-to-time encoding. These results are also visible in the scores of the ICE-T questionnaire [46], where NL+JP and NL+ANC are quite closely ranked in terms of *Essence*, that is the capability of the visualization of communicating both overview and context of the data (see Figure 10 and Table 4)

6.6 Limitations

In our second study, we explored how the scale of the graph (in terms of the number of nodes and edges as well as timeslices) affected the ability of the participants to extract insights and generate knowledge from a real-world dataset. We had to face space issues with JP, as beyond 8 timeslices there is not nearly enough screen space to depict the graph in its entirety in all timeslices while keeping a sufficient zoom level to keep labels and nodes readable.

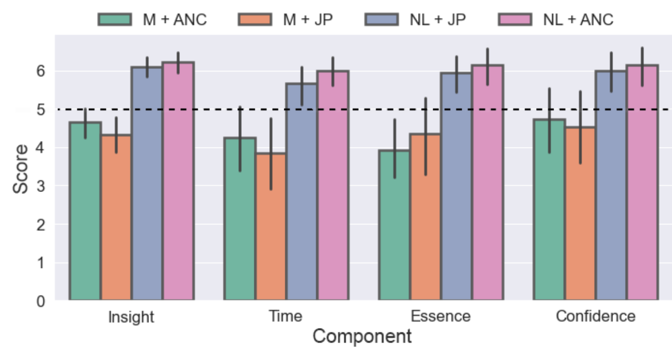

Fig. 11: Results for each of the ICE-T heuristics. The bars each represent one of the techniques being evaluated. The scores vary from 1 (Strongly Disagree) to 7 (Strongly Agree). The dashed line indicates a score of 5, which is considered the minimum .

While the results of the evaluation are clear in showing the participants' preference, two elements might have negatively affected M-based representations leading to a lower score. These should be considered when elaborating on the study results, and, therefore, we discuss them as limitations.

First, M visualizations appeared to be sparse ("[There was] *lots of whitespace*"). Co-authorship networks appear as rapidly changing cliques and we picked the biggest cliques in order to have a sufficient amount of nodes and edges. However, we soon hit an upper bound on the number of nodes, as space requirements for M are particularly demanding to accommodate node labels. While it is commonly accepted that sparse graphs are better visualized as NL diagrams, we were not testing task performance, but rather the expressive power of combinations of structural and temporal encodings in a dynamic scenario. Nonetheless, it remains untested whether M-based approaches would provide improved task performance compared to NL in a dynamic scenario with larger, denser graphs.

Second, one participant observed that there could be some negative bias towards M-based approaches as NL are more common. We did not measure or counteract bias effects.

Finally, we remark that this study provides empirical evidence of the expressive power and preference of different dynamic network visualizations in a more realistic scenario, which includes larger graphs and more nuanced interactions. However, we can't conclude whether the findings of the first study apply to larger graphs, which would require further studies in this direction.

7 Takeaways

Our results suggest that, within the afforementioned limitations, the selection of NL as a structural representation leads to better performance, and participants generally prefer NL over M. In our first experiment, M was closer to NL regarding performance and preference, considering the simple interactions and few timeslices. In the second study, a broader gap in preference appeared between the two. Existing research proved the potential of M over NL representations for specific tasks in a static context (see, e.g., [11], [15]). Despite offering interactions and reordering techniques for M-based representations, these need to be adapted and extended for a dynamic context in order to emphasize the topological structures that exist and change over time. This has two interpretations.

First, there is an intrinsic difficulty in reporting the dynamics of a graph using matrices: as all the rows/columns are visible simultaneously, choices have to be made on whether to hide the currently inactive ones. To help maintain the mental map, we chose to show all rows and columns for all timeslices. This enabled the participants to orient themselves but simultaneously created clutter that played against them rather than providing a more comprehensive view of the dataset and its temporal evolution. The first study's results could also support this: M could provide a similar performance to NL as only edges were added/removed.

Second, M representations are undoubtedly oriented towards a more expert audience of users. In the first study, some participants (students, see Section 5) commented that they needed to redraw the M representation as a NL diagram. Whereas in the second study, even expert participants admitted a potential bias as they were used to interacting with NL diagrams compared to M representations in their daily work.

Within the context of our two studies and the basic implementations of the graph representations, temporal encodings, and available interactions these results appear in contrast to the ones we obtained in our exploratory study [27]. We would like to remark that our previous work was intended as an *exploratory* study to gather empirical evidence about the problem to be further evaluated in this work. In fact, we focused on a significantly reduced set of tasks, used smaller graphs (both number of nodes/edge and timeslices), did not consider mental map preservation, and neither present a statistical analysis of the results. Therefore, we believe that the *combination* of the findings of the two study can support users in making the best informed decisions for the task at hand. The insights, quotes, and results of our exploratory study [27] should be mediated with the results of both experiments in this work.

The results of both studies highlight that it is easier for users to transition to a dynamic network exploration scenario with NL representations. At the same time, for M there are still open, underinvestigated problems related to the presentation of the temporal dynamics. Furthermore, when considering the use of M for dynamic networks, the target user groups' expertise should be considered.

Concerning the temporal encodings, we found AN, ANC, and JP to be the best performing and most preferred across our two experiments. In the first study, AN generally had lower response times and higher accuracy than JP. AN was closely followed by ANC, which was the most preferred due to the presence of the manual time slider. JP showed evident limitations when we scaled up the number of timeslices, but it is still the encoding of choice for overview tasks. Participants remarked that the eye needed to travel some distance and locate the same position within the graph to focus on it before they could see what is happening to an individual node or edge. This hints that JP may not be as well suited for low-level tasks as ANC but was regarded as a suitable approach for getting an overview of the entire network and how it changes across all timeslices. Nonetheless, the space requirements of JP should be carefully considered when designing a dynamic network visualization system.

8 Conclusion

In this paper, we investigated the design space of dynamic network visualization along its two major dimensions, structural representation and temporal encoding, considering the effects of simpler and more advanced interaction techniques. We presented two studies: Firstly, we conducted a user study, assessing response times, accuracy, and preferences. We evaluated the results against our research hypotheses through a complete statistical analysis. Secondly, based on these results, we selected the best-performing structural and temporal encodings and improved interactions according to participants' feedback. The resulting visualizations were compared in a heuristic evaluation [46] aimed at investigating to which extent they support conducting exploratory analysis and insight generation.

In accordance with the results of the first study, participants favored NL over M for the structural encoding and preferred ANC over JP for the temporal encoding. We condensed our results into a series of takeaways and discussed the limitations of our studies, which opened several interesting future research directions. First, we believe the tradeoff between layout quality and stability [29] should be further investigated in this context for both NL and M. Second, while both ANC and JP had high scores for NL diagrams, it would be interesting to see if and how both approaches could be combined interactively to provide an overall better experience for network exploration and analysis. Finally, we believe that the staged animation technique proposed by Bach et al. [75] could be applied to both NL and M helping users orient when transitioning from one timeslice to the other, with visual cues indicating the location and type of change (appearance or disappearance).

Acknowledgments

This work was conducted within the framework of the ArtVis (P35767) project funded by the Austrian Science Fund (FWF).

REFERENCES

[1] W. T. Tutte, "How to draw a graph," *Proceedings of the London Mathematical Society*, vol. 3, no. 1, pp. 743–767, 1963.

[2] I. Herman, G. Melancon, and M. Marshall, "Graph Visualization and Navigation in Information Visualization: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[3] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "A Taxonomy and Survey of Dynamic Graph Visualization," *Computer Graphics Forum*, vol. 36, no. 1, p. 133–159, 2017.

[4] S. Hadlak, H. Schumann, and H.-J. Schulz, "A Survey of Multi-faceted Graph Visualization," in *Eurographics Conference on Visualization*. The Eurographics Association, 2015.

[5] N. Kerracher, J. Kennedy, and K. Chalmers, "The Design Space of Temporal Graph Visualisation," in *EuroVis - Short Papers*. The Eurographics Association, 2014.

[6] R. Keller, C. M. Eckert, and P. J. Clarkson, "Matrices or Node-Link Diagrams: Which Visual Representation is Better for Visualising Connectivity Models?" *Information Visualization*, vol. 5, no. 1, pp. 62–76, 2006.

[7] M. Okoe, R. Jianu, and S. Kobourov, "Node-Link or Adjacency Matrices: Old Question, New Insights," *IEEE Trans. on Visualization and Computer Graphics*, vol. 25, no. 10, pp. 2940–2952, 2019.

[8] N. Henry and J.-D. Fekete, "Matlink: Enhanced matrix visualization for analyzing social networks," in *IFIP Conference on Human-Computer Interaction*. Springer, 2007, pp. 288–302.

[9] S. Rufiange and G. Melançon, "Animatrix: A matrix-based visualization of software evolution," in *2014 second IEEE working conference on software visualization*. IEEE, 2014, pp. 137–146.

[10] D. Archambault and H. C. Purchase, "The "map" in the mental map: Experimental results in dynamic graph drawing," *International Journal of Human-Computer Studies*, vol. 71, no. 11, pp. 1044–1055, 2013.

[11] M. Okoe, R. Jianu, and S. Kobourov, "Revisited Experimental Comparison of Node-Link and Matrix Representations," in *Graph Drawing & Network Visualization*. Springer International Publishing, 2018, pp. 287–302.

[12] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch, "The Aesthetics of Graph Visualization," in *Proc. of the Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*. The Eurographics Association, 2007, p. 57–64.

[13] E. D. Giacomo, W. Didimo, G. Liotta, F. Montecchiani, and A. Tappini, "Comparative Study and Evaluation of Hybrid Visualizations of Graphs," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2022.

[14] D. Ren, L. R. Marusich, J. O'Donovan, J. Z. Bakdash, J. A. Schaffer, D. N. Cassenti, S. E. Kase, H. E. Roy, W.-y. S. Lin, and T. Höllerer, "Understanding node-link and matrix visualizations of networks: A large-scale online experiment," *Network Science*, vol. 7, no. 2, p. 242–264, 2019.

[15] M. Ghoniem, J. . Fekete, and P. Castagliola, "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations," in *IEEE Sym. on Information Visualization*, 2004, pp. 17–24.

[16] H. C. Purchase, "The Effects of Graph Layout," in *Proc. of the Australasian Conference on Computer Human Interaction*. IEEE Computer Society, 1998, p. 80.

[17] H. Purchase, D. Carrington, and J.-A. Allder, "Empirical Evaluation of Aesthetics-based Graph Layout," *Empirical Software Engineering*, vol. 7, 2002.

[18] M. Abdelaal, N. D. Schiele, K. Angerbauer, K. Kurzhals, M. Sedlmair, and D. Weiskopf, "Comparative Evaluation of Bipartite, Node-Link, and Matrix-Based Network Representations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 896–906, 2023.

[19] D. Archambault, H. Purchase, and B. Pinaud, "Animation, Small Multiples, and the Effect of Mental Map Preservation in Dynamic Graphs," *IEEE Trans. on Visualization and Computer Graphics*, 2011.

[20] D. Archambault and H. C. Purchase, "Mental Map Preservation Helps User Orientation in Dynamic Graphs," in *Graph Drawing & Network Visualization*. Springer, 2013, pp. 475–486.

[21] ——, "Can Animation Support the Visualisation of Dynamic Graphs?" *Information Sciences*, p. 495–509, 2016.

[22] H. C. Purchase, E. Hoggan, and C. Görg, "How important is the "mental map"?–an empirical investigation of a dynamic graph layout algorithm," in *International Symposium on Graph Drawing*. Springer, 2006, pp. 184–195.

[23] S. Ghani, N. Elmqvist, and J. S. Yi, "Perception of Animated Node-Link Diagrams for Dynamic Graphs," *Computer Graphics Forum*, 2012.

[24] C. D. Linhares, J. R. Ponciano, J. G. S. Paiva, B. A. Travençolo, and L. E. Rocha, "A comparative analysis for visualizing the temporal evolution of contact networks: a user study," *Journal of Visualization*, vol. 24, no. 5, pp. 1011–1031, 2021.

[25] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk, "Dynamic network visualization withextended massive sequence views," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 8, pp. 1087–1099, 2013.

[26] C. D. Linhares, B. A. Travençolo, J. G. S. Paiva, and L. E. Rocha, "DyNetVis: a system for visualization of dynamic networks," in *Proceedings of the symposium on applied computing*, 2017, pp. 187–194.

[27] V. Filipov, A. Arleo, and S. Miksch, "Exploratory User Study on Graph Temporal Encodings," in *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, 2021, pp. 131–135.

[28] M. Bostock, V. Ogievetsky, and J. Heer, "D3 Data-Driven Documents," *IEEE Trans. on Visualization and Computer Graphics*, vol. 17, no. 12, p. 2301–2309, 2011.

[29] U. Brandes and M. Mader, "A quantitative comparison of stress-minimization approaches for offline dynamic graph drawing," in *International Symposium on Graph Drawing*. Springer, 2011, pp. 99–110.

[30] S. Diehl and C. Görg, "Graphs, They Are Changing," in *Graph Drawing*. Springer Berlin Heidelberg, 2002, pp. 23–31.

[31] A. Arleo, S. Miksch, and D. Archambault, "Event-based Dynamic Graph Drawing without the Agonizing Pain," in *Computer Graphics Forum*. Wiley Online Library, 2022.

[32] M. Baur, M. Benkert, U. Brandes, S. Cornelsen, M. Gaertler, B. Köpf, J. Lerner, and D. Wagner, "Visone software for visual social network analysis," in *International Symposium on Graph Drawing*. Springer, 2001, pp. 463–464.

[33] U. Brandes and S. R. Corman, "Visual unrolling of network evolution and the analysis of dynamic discourse," *Information Visualization*, vol. 2, no. 1, pp. 40–50, 2003.

[34] C. Collberg, S. Kobourov, J. Nagra, J. Pitts, and K. Wampler, "A system for graph-based visualization of the evolution of software," in *Proceedings of the 2003 ACM symposium on Software visualization*, 2003, pp. 77–ff.

[35] T. Crnovrsanin, J. Chu, and K.-L. Ma, "An Incremental Layout Method for Visualizing Online Dynamic Graphs." *J. Graph Algorithms Appl.*, vol. 21, no. 1, pp. 55–80, 2017.

[36] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee, "GraphAEL: Graph animations with evolving layouts," in *International Symposium on Graph Drawing*. Springer, 2003, pp. 98–110.

[37] ——, "Exploring the computing literature using temporal graph visualization," in *Visualization and Data Analysis 2004*, vol. 5295. SPIE, 2004, pp. 45–56.

[38] P. Simonetto, D. Archambault, and S. Kobourov, "Event-based dynamic graph visualisation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 7, pp. 2373–2386, 2018.

[39] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, and J.-D. Fekete, "Matrix reordering methods for table and network visualization," in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 693–716.

[40] S. Miksch and W. Aigner, "A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data," *Computers & Graphics*, vol. 38, pp. 286–290, 2014.

[41] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual Comparison for Information Visualization," *Information Visualization SAGE*, vol. 10, no. 4, p. 289–309, 2011.

[42] W. Javed and N. Elmqvist, "Exploring the design space of composite visualization," in *IEEE Pacific Visualization Symposium*, 2012, pp. 1–8.

[43] Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, A. Pedro, Sciaini, Marco, Scherer, and Cédric, *viridis - Colorblind-Friendly Color Maps for R*, r package version 0.6.2.

[44] E. R. Tufte, "The Visual Display of Quantitative Information," *Quantitative Information*, p. 13, 1983.

[45] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 443–456, 2005.

[46] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko, "A Heuristic Approach to Value-Driven Evaluation

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3310019

of Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 491–500, 2019.

[47] J.-w. Ahn, C. Plaisant, and B. Shneiderman, "A Task Taxonomy for Network Evolution Analysis," *IEEE Trans. on Visualization and Computer Graphics*, vol. 20, no. 3, p. 365–376, 2014.

[48] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the international AAAI conference on web and social media*, vol. 3, no. 1, 2009, pp. 361–362.

[49] B. Bollobás, C. Borgs, J. T. Chayes, and O. Riordan, "Directed scale-free graphs." in *SODA*, vol. 3, 2003, pp. 132–139.

[50] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[51] A. Hagberg and D. Conway. (2020) NetworkX: Network Analysis with Python.

[52] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[53] V. Yoghourdjian, D. Archambault, S. Diehl, T. Dwyer, K. Klein, H. C. Purchase, and H.-Y. Wu, "Exploring the limits of complexity: A survey of empirical studies on graph visualisation," *Visual Informatics*, vol. 2, no. 4, pp. 264–282, 2018.

[54] SurveyJS - Survey and Form JavaScript Libraries. https://surveyjs.io/. [Online; accessed: August 2022].

[55] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, pp. 3–62, 1936.

[56] V. Filipov, A. Arleo, M. Bögl, and S. Miksch, "On Time and Space: An Experimental Study on Graph Structural and Temporal Encodings," 2022. [Online]. Available: https://arxiv.org/abs/2208.13716

[57] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[58] R. Vallat, "Pingouin: statistics in Python," *Journal of Open Source Software*, vol. 3, p. 1026, 2018.

[59] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[60] K. Backhaus, B. Erichson, and R. Weiber, *Fortgeschrittene multivariate Analysemethoden: eine anwendungsorientierte Einführung*. Springer-Verlag, 2015.

[61] J. Bortz, *Statistik: Für Sozialwissenschaftler*. Springer-Verlag, 2013.

[62] J. Hedderich and L. Sachs, *Angewandte Statistik*. Springer, 2016.

[63] C. Weiß and C. Weiß, *Basiswissen medizinische statistik*. Springer, 2005, vol. 4.

[64] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[65] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[66] C. A. Boneau, "The effects of violations of assumptions underlying the t test." *Psychological bulletin*, vol. 57, no. 1, p. 49, 1960.

[67] H. O. Posten, "Robustness of the two-sample t-test," in *Robustness of statistical methods and nonparametric statistics*. Springer, 1984, pp. 92–99.

[68] E. Schminder, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, "Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution," *Methodology – European Research Journal of Methods for the Behavioral and Social Sciences*, vol. 6, no. 4, pp. 147–151, 2010.

[69] J.-D. Fekete, "Reorder. js: A javascript library to reorder tables and networks," in *IEEE VIS 2015*, 2015.

[70] Citevis citation datafile. http://www.cc.gatech.edu/gvu/ii/citevis/infovis-citation-data.txt. [Online; accessed February 2023].

[71] Zoom. https://zoom.us/. [Online; accessed June 2023].

[72] Angular. https://angular.io/. [Online; accessed June 2023].

[73] d3js. https://d3js.org/. [Online; accessed June 2023].

[74] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Visualization of Time-Oriented Data*, 2nd ed. Springer Publishing Company, Incorporated, forthcoming, 2023.

[75] B. Bach, E. Pietriga, and J.-D. Fekete, "Visualizing Dynamic Networks with Matrix Cubes," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2014, p. 877–886.

Velitchko Filipov is a PhD student in the Visual Analytics research unit (CVAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. His research interests include information visualization and visual analytics of dynamic graphs and networks. He is a Graduate Student Member of the IEEE.

Alessio Arleo is a Post-Doc in the Visual Analytics research unit (CVAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. He obtained a PhD in 2018 at University of Perugia, Italy (thesis: Distributed Large Graph Visualization: Algorithms and Experiments). He holds a BSc in Computer and Electronic engineering and MSc in Computer and TLC engineering. His research interests include network visualization, distributed algorithms, and software engineering.

Markus Bögl is a Post-Doc in the Visual Analytics research unit (CVAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. During his PhD (thesis: Visual Analysis of Periodic Time Series Data, finished in 2020), he researched visual analytics for time series analysis, combining visualization, interaction, and statistical analysis. His research interest includes visual analytics for time-oriented data, space and spatiotemporal data, data quality issues (uncertainty and outlier detection), and parametrizations of models for segmenting and labeling of time series data.

Silvia Miksch is University Professor and head of the Visual Analytics research unit (CVAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. She served as paper co-chair of several conferences including IEEE VAST 2010, 2011, 2020, and VIS Overall Papers Chair (IEEE VIS 2021) as well as EuroVis 2012 and on the editorial board of several journals including IEEE TVCG and Computer Graphics Forum. She acts/acted in various strategic committees, such as the chair of the EuroVis steering committee and the VIS Executive Committee. In 2020 she was inducted into The IEEE Visualization Academy. Her main research interests are Visualization/Visual Analytics (particularly Focus+Context and Interaction), Space, and Time.