Journal Title
XX(X):1–??
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Seeing Graphs Like Humans: Benchmarking Computational Measures and MLLMs for Similarity Assessment

Seokweon Jung[1,2]* Jeongmin Rhee[3], Seoyoung Doh[4], Hyeon Jeon[2], Ghulam Jilani Quadri[5] and Jinwook Seo[2]

Abstract

Comparing graphs to identify similarities is a fundamental task in visual analytics of graph data. To support this, visual analytics systems frequently employ quantitative computational measures to provide automated guidance. However, it remains unclear how well these measures align with subjective human visual perception, thereby offering recommendations that conflict with analysts' intuitive judgments, potentially leading to confusion rather than reducing cognitive load. Multimodal Large Language Models (MLLMs), capable of visually interpreting graphs and explaining their reasoning in natural language, have emerged as a potential alternative to address this challenge. This paper bridges the gap between human and machine assessment of graph similarity through three interconnected experiments using a dataset of 1,881 node-link diagrams. Experiment 1 collects relative similarity judgments and rationales from 32 human participants, revealing consensus on graph similarity while prioritizing global shapes and edge densities over exact topological details. Experiment 2 benchmarks 16 computational measures against these human judgments, identifying Portrait divergence as the best-performing metric, though with only moderate alignment. Experiment 3 evaluates the potential of three state-of-the-art MLLMs (GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5) as perceptual proxies. The results demonstrate that MLLMs, particularly GPT-5, significantly outperform traditional measures in aligning with human graph similarity perception and provide interpretable rationales for their decisions, whereas Claude Sonnet 4.5 shows the best computational efficiency. Our findings suggest that MLLMs hold significant promise not only as effective, explainable proxies for human perception but also as intelligent guides that can uncover subtle nuances that might be overlooked by human analysts in visual analytics systems.

Keywords

Graph visualization, graph similarity, graph comparison, human perception, multimodal large language model.

Introduction

Comparing multiple graphs visually to discern structural variations and topological shifts is a fundamental task in visual analytics across diverse domains[1–3], enabling analysts to discern dominant trends or outlying patterns. This is particularly critical in dynamic graph analysis[4], where understanding temporal evolution commonly involves slicing dynamic graphs into multiple static snapshots and assessing their similarities and differences over time[5,6]. To support this, modern visual analytics systems typically employ a dual approach: leveraging human visual perception for qualitative interpretation while utilizing quantitative computational measures to reduce cognitive load and guide the user's attention[7–10].

While this synergistic approach aims to mitigate the complexity of analysis, its foundational premises remain insufficiently validated in graph comparison. In other domains, such as scatterplot analysis, the alignment between human perception and computational metrics has been extensively studied to optimize system design[11,12]. In contrast, although previous studies have proposed design guidelines for comparison tasks[13,14], the extent to which humans can effectively perceive differences between graphs remains underexplored[15]. Consequently, to design effective visual analytics systems, it is imperative to first establish a baseline understanding of the environments and support mechanisms that facilitate human graph comparison.

Graph comparison is compounded by diverse properties such as size, density, and layout[16,17], alongside critical factors like visual scalability and mental map preservation[18,19]. This complexity makes it unclear which graph features or visual cues humans prioritize when assessing similarity, and

[1]LLM Innovation Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
[2]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea
[4]College of Liberal Studies, Seoul National University, Seoul, Republic of Korea
[5]School of Computer Science, University of Oklahoma, Norman, Oklahoma, United States

Corresponding author:
Jinwook Seo, HCI Lab, Department of Computer Science and Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
Email: jseo@snu.ac.kr

whether these perceptual judgments are consistent across different observers. Therefore, we propose our first research question: *Can humans discern similarity and differences between graph visualizations, and which factors significantly affect these decisions?*

To complement human judgment, computational measures offer objective quantification. While a wide range of metrics exists, from classical graph edit distances to advanced methods such as spectral distances[20–22], their alignment with human visual perception remains largely unverified. This lack of empirical validation poses a critical risk: automated systems may offer recommendations that conflict with analysts' intuitive judgments, potentially leading to confusion and mistrust rather than reducing cognitive load[23–25]. This knowledge gap leads to our second research question: *Do traditional computational measures capture the similarities and differences that humans perceive visually?*

Seeking a computational method that better aligns with human intuition, we examine methodologies capable of interpreting the holistic and semantic context of graph visualizations. Recent advances in Multimodal Large Language Models (MLLMs) have emerged as a computational method for supporting visual analytics tasks by demonstrating notable capabilities in interpreting visual content, including charts and diagrams[26–28]. Although some research suggests that the MLLMs are nearly blind at comprehending complex visualizations[29], recent studies indicate they are overcoming these barriers, demonstrating promising capabilities in layout generation and basic perception[30–32]. Despite these advancements, their utility in graph comparison remains underexplored, yet their evolving visual understanding suggests they hold significant potential to act as proxies for human perception[33]. Based on this potential, we propose our third research question: *Do MLLMs possess the capability to align with human perception and guide users in graph comparison tasks?*

To address these three research questions, we conduct a series of three interconnected experiments. To validate our questions under diverse conditions, we create a comprehensive dataset comprising 1,881 node-link diagrams. These are generated from both synthetic and real-world data, systematically varying in size, density, and layout, which are graph attributes known to have great influence on human perception of graph visualizations[34,35].

In Experiment 1, to account for the subjective nature of human similarity perception, we employ a relative-comparison task in which participants judge which of two target graphs is more similar to a reference graph[36,37]. We collect 2,208 relative similarity judgments, confidence ratings, and corresponding explanations from 32 participants. Our analysis reveals consistent trends in human perception of graph similarity across varying graph conditions, providing insights into the factors influencing both perception and self-assessed certainty.

In Experiment 2, we evaluate the extent to which computational graph similarity measures align with human perceptual judgments. We select 16 established measures representing a range of theoretical backgrounds based on a comprehensive literature review[22,38–42]. We assess agreement between human choices and computational scores, identify which measures most closely reflect human judgments, and analyze how discrepancies relate to participant confidence.

Finally, in Experiment 3, we explore the potential of three state-of-the-art MLLMs, GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5, to visually assess graph similarity. The models are prompted to perform the same relative comparison task as human participants, including providing confidence ratings and explanations for their decisions. Our analysis demonstrates that MLLMs exhibit a higher overall agreement with human judgments than traditional computational measures. Furthermore, they provide interpretable explanations for their decision criteria, suggesting they can effectively lower the barrier to visual analytics of graph data.

Synthesizing the insights from this series of experiments, we propose evidence-based guidelines for designing visual analytics systems that include graph comparison tasks, such as observing temporal evolution or detecting anomalies in dynamic graphs. We outline how to effectively structure graph comparison tasks for human analysts, select perception-aligned computational metrics, and strategically leverage MLLMs to provide explainable guidance, ultimately aiming to bridge the gap between human intuition and machine quantification.

In summary, this work makes the following contributions:

- We construct a large-scale dataset of graph visualizations systematically varied by topological and visual properties to benchmark similarity perception.
- We provide empirical evidence of consistent patterns in human perception of graph similarity and identify key factors influencing these judgments.
- We identify the specific computational similarity measures that best approximate human visual perception of graph differences.
- We demonstrate that state-of-the-art MLLMs closely align with human perception and provide interpretable rationales, highlighting their potential as effective assistants in visual analytics.

Related Work

We discuss relevant literature on computational measures for graph comparison as well as visualization techniques for comparative analysis tasks involving graphs.

Computational Graph Comparison

Measure-based Comparison Graphs can be quantitatively compared using graph similarity measures, often referred to as graph distances. Various graph distance measures have been developed to facilitate graph comparisons. In our study, we review literature published in SCI(E) journals and presented at leading international conferences over the past decade. Our investigation included surveys on graph and network similarity/distance that compared multiple measures[22,38,39], as well as studies that evaluated and contrasted various similarity measures to provide further research guidance[40–42]. This literature review enables us to both classify the existing graph similarity measures and identify those applicable to our target graphs.

A key classification criterion, as highlighted in the survey studies, is node correspondence[22,38,39]. When graphs

contain node labels and a one-to-one correspondence exists between nodes, this information is critical for node-based comparisons. Measures such as graph edit distance[20] and DELTACON[43] are designed for scenarios with known node correspondence (KNC). However, as noted in our study design, our target graphs are characterized by unknown node correspondence (UNC), thus we exclude KNC-based methods from our analysis.

UNC graph similarity measures determine similarity based on high-level structural features rather than relying on node-specific details. Depending on the granularity of the features considered, these measures are typically categorized with the levels of granularity: local and global[22,38,39].

The applicability and performance of these measures vary significantly depending on the specific characteristics of the graphs, such as whether they are directed, weighted, or contain self-loops[42]. Moreover, frameworks have categorized these similarity measures based on their scope, distinguishing them as micro-level, meso-level, and macro-level measures[39]. Several studies extensively reviewed and compared these diverse graph similarity metrics to better understand their advantages, limitations, and practical applicability[40–42].

However, despite extensive comparative analyses of computational measures, the literature currently lacks empirical studies investigating how these measures align with human perceptions when the results are visually presented. To address this gap, we select representative computational measures suited to our experimental context. These measures span diverse theoretical backgrounds[38], ranging from straightforward approaches such as Jaccard distance and edit distance[20] to more sophisticated spectral similarity measures[44], graph kernels[45], and graphlet-based techniques[46].

Human Visual Comparison Another important method for comparing graphs is visual analysis, which helps users directly identify similarities and differences through graphical representations[47]. Effective visual comparison requires careful consideration of visualization principles and perceptual guidelines[13,14].

Visual comparison methods are particularly prevalent in analyzing dynamic graphs, which frequently represent complex and large-scale temporal datasets[4,5]. Taxonomies of dynamic graph inspection explicitly include comparative analysis tasks[48]. Graph visualization techniques are extensively employed to present both overall patterns[7–9] and intricate details[8,9,49,50] of graph data.

The comparison itself is recognized as a foundational, low-level visual analytic activity[51] involving the examination of attributes and relationships within data. Previous studies have investigated the perceptual aspects of graph visualizations, typically focusing on the readability or performance of different visualization techniques[15,52–54]. More recently, research has examined how different node-link layouts influence the perception of graph properties[35]. However, studies explicitly investigating human perception of graph similarity remain limited.

Empirical studies using graph data have highlighted substantial variability in perceptual performance and preferences according to graph size and complexity[22,55]. One notable study incrementally adjusted simple graph attributes to evaluate perceptual thresholds for detecting graph differences[56]. Although there have been empirical examinations of perceptual differences related to specific graph features[35,57], none have explicitly targeted similarity perception. Research examining temporal network changes has primarily focused on mental map preservation rather than similarity[19]. In our study, we quantitatively investigate the relationship between human visual perception and computational graph similarity measures.

Quantifying Visual Perception

Quantifying human perception, which is an inherently subjective phenomenon, has been explored across various visualization contexts[13,14]. Researchers have developed methods to measure perceptual judgments in different visualization domains, including visual encoding[58,59], clustering[36,60,61], and time-series data[37].

While perceptual tasks involving node-link diagrams and adjacency matrices have also been studied[62], specific research on human perception of graph similarity is notably scarce. Existing related work primarily focused on the perceptual effects of varying graph properties or visualization techniques rather than explicitly addressing similarity perception[35]. Our research aims to fill this gap by explicitly measuring perceptual similarity judgments in graph visualizations.

MLLMs for Graph Visualization

The emergence of powerful Multimodal Large Language Models (LLMs) has enabled the interpretation and generation of visual content through natural language descriptions[63–66]. While some studies have characterized MLLMs as being nearly "blind" in their visualization interpretation capabilities due to early limitations[29], their potential remains substantial, and further advancements are expected to significantly impact future visualization research[33]. A growing body of work demonstrates that state-of-the-art models are overcoming these initial barriers, exhibiting significant potential in both generative and perceptual tasks. For instance, they have been utilized to create visualization guidelines[28,67,68], evaluate visualization effectiveness[27], and generate novel visualizations[26].

Specifically in the domain of graph visualization, although the inherent complexity of graph topology poses a unique barrier, recent improvements in model performance have spurred research into leveraging MLLMs for diverse graph-related tasks. Di Bartolomeo et al. explored the generative capabilities of ChatGPT for graph drawing, revealing that while challenges remain in handling complex constraints, the model exhibits significant potential in applying layout algorithms and coordinating spatial arrangements via text prompts[30]. Building on this, Fan et al. conducted a comprehensive evaluation of LLMs on graph layout tasks, reporting that advanced models, such as GPT-4, can effectively adhere to aesthetic constraints and optimize layout metrics, particularly for small-scale graphs[31]. Shifting the focus from generation to perception, Miller et al. investigated MLLMs' understanding of network visualization principles[32]. Their findings suggest that these

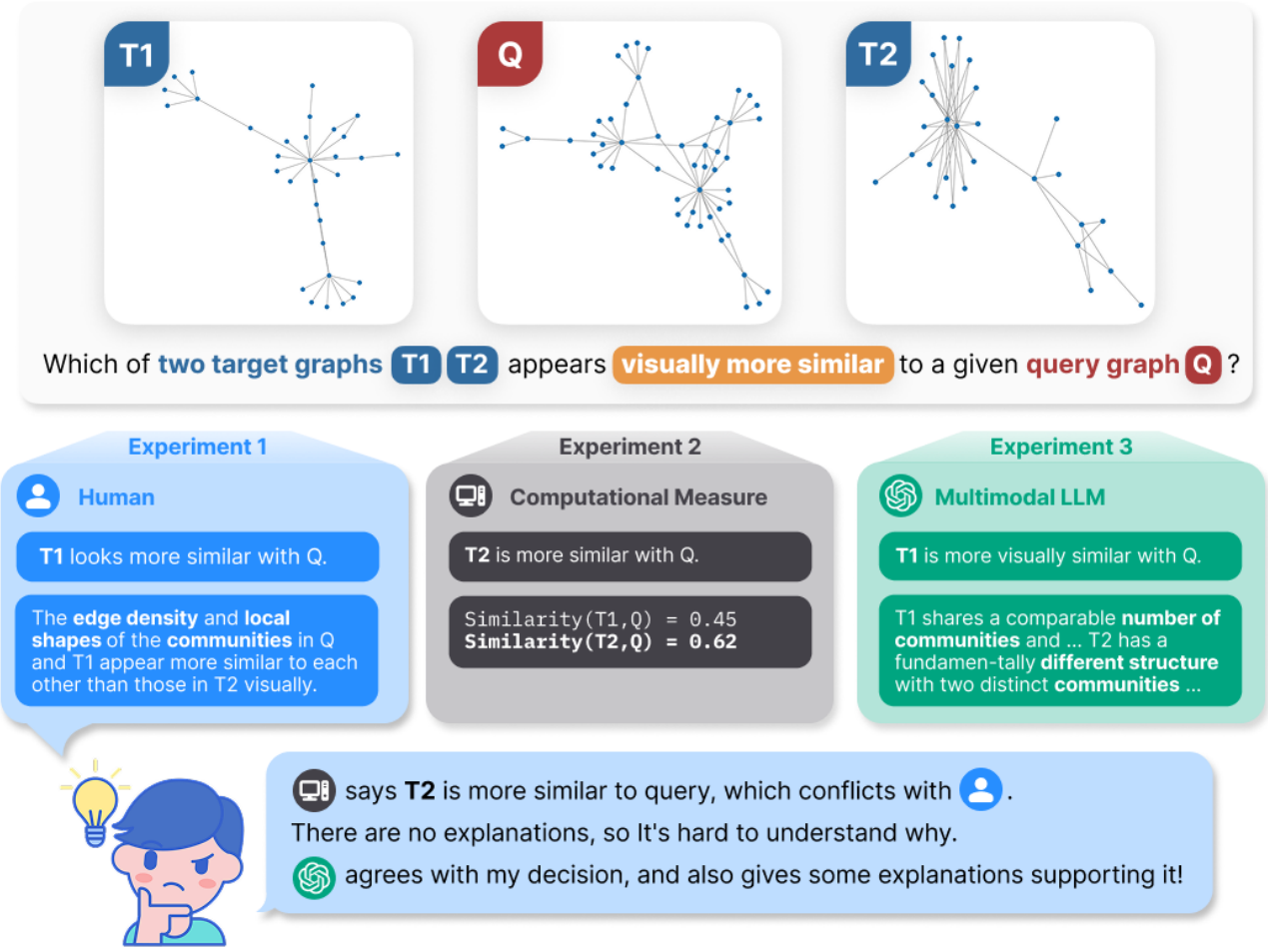

Figure 1. An overview of our research methodology, structured into three interconnected experiments designed to investigate graph comparison capabilities in humans and machines. Experiment 1 assesses human competence in graph similarity assessment through indirect similarity measurements derived from visual perception (RQ1). Experiment 2 computes pairwise similarities using 16 distinct computational graph similarity measures and compares them with human decisions to determine the alignment between humans and machines (RQ2). Experiment 3 evaluates the relative graph similarity assessment capabilities of MLLMs, analyzing both their alignment with human perception and the interpretability of their reasoning (RQ3). Our results demonstrate that MLLMs exhibit a higher alignment with human judgment than computational measures, qualifying them as superior perceptual proxies. Furthermore, by providing interpretable decision rationales, they serve as a more effective method for assisting human analysts in graph comparison tasks.

models are beginning to align with human judgments regarding fundamental design aspects, such as visual clutter and Gestalt principles.

Despite these significant advancements in layout generation and single-graph assessment, the specific task of graph comparison remains underexplored in the existing literature. Comparing graphs is a foundational analytic activity, essential not only for distinguishing structural differences between static graphs but also for analyzing temporal evolution and detecting anomalies in dynamic graph analysis. In our study, we explore the extent to which current MLLMs comprehend graph similarity, investigate their decision criteria, and evaluate the consistency of their assessments relative to human judgments and computational measures.

Methodology

We conduct a series of three interconnected experiments to provide empirical insights for designing effective visual analytics systems in which users must compare graphs across multiple graph visualizations (see Figure 1). Our study investigates three primary research questions:

- **RQ1:** Can humans discern similarity and differences between graph visualizations, and which factors significantly affect these decisions?
- **RQ2:** Do traditional computational measures capture the similarities and differences that humans perceive visually?
- **RQ3:** Do MLLMs possess the capability to align with human perception and guide users in graph comparison tasks?

To address these questions systematically, we first construct a comprehensive dataset of graph visualizations. Subsequently, we systematically control graph size, edge density, and layout algorithms, employing them as independent variables known to influence empirical perception [16,35,57].

Experiment 1 then uses this dataset to assess how human participants perceive similarities between graphs (**RQ1**). To mitigate individual differences in perceptual sensitivity

Table 1. Three independent variables. We utilize the notion of linear edge density ($\frac{|E|}{|V|}$) to follow the classification criteria by Yoghourdjian et al.[16]. Layout algorithms are selected based on popularity and ease of use[35].

Var.	Description	Value range
Size (***N***)	Number of nodes	small:[10, 20], medium:[21, 50], large:[51, 200], very large:[201, 400]
Density (***D***)	Number of edges divided by nodes	sparse:[1, 2), dense:[2, 3), very dense:[3, 10]
Layout (***L***)	Algorithm to place nodes and edges	force-directed (F-R), circular, multidimensional scaling (UMAP)

and the inherent difficulty of quantifying similarity on an absolute scale, we employ a relative comparison task[37]. We develop a web-based experimental interface and collect a total of 2,208 responses from 32 participants. The subsequent Experiment 2 and Experiment 3 focus on identifying which computational methods best align with the human similarity assessments collected in Experiment 1 **(RQ2, RQ3)**. We view this alignment process as the evaluation of the accuracy of computational measures that proxy human perception, adopting a data-driven approach to evaluate Visual Quality Measures (VQMs) based on perceptual data[69].

Study Design

Target Graph Specification To establish robust foundations and provide reliable ground-truth data, our research systematically examines perceived graph similarity using the most fundamental form of graph visualization. We focus on undirected, unweighted graphs without self-loops. Despite their simplicity, these graphs are standard in graph visualization research, and many visualization techniques[49,70] and similarity measures[71,72] are explicitly designed for such configurations.

Another critical consideration in graph comparison tasks is node correspondence, which can be known node correspondence (KNC) or unknown node correspondence (UNC)[38]. When node correspondence is known, visualizations typically include node-specific information through additional visual channels, such as color, to represent node attributes. Since our goal is to study basic graph visualizations without the influence of additional node-specific information, we focus on graphs with UNC, thereby avoiding visual cues that might bias assessments.

Moreover, evaluating disconnected graphs requires comparing multiple graphs simultaneously, which raises the difficulty of the comparison task. To maintain consistency and simplicity, we limit our investigation to graphs consisting of a single connected component.

Stimuli Generation We generate 1,881 node-link diagrams comprising 1,152 synthetic graphs and 729 real-world graphs. Guided by a survey of empirical studies on graph visualization[16] and research investigating the factors influencing node-link diagrams[17,57], the dataset covers a broad spectrum of graph sizes, densities, and layouts. The following subsections detail the generation process.

Data Sources Synthetic graphs offer diverse examples, while real-world graphs provide practical relevance[56]. Thus, we include both data sources for our experiment. Synthetic graphs are generated using four different graph generation algorithms, all of which generate graphs with distinct characteristics (Figure 2).

- **GNM**: Graphs with randomly generated M edges between N nodes. While the related study[57] employs the Erdős–Rényi model[76], we utilized GNM to directly control the edge density.
- **BBA (Barabási–Albert)**: Graphs with power-law degree distributions by connecting edges between new nodes to existing nodes with high degree[73].
- **NWS (Newman–Watts–Strogatz)**: Graphs with a ring over nodes and connection between their k nearest neighbors[74].
- **SBM (Stochastic Block Model)**: Graphs partitioned into blocks of arbitrary sizes and edges are placed between pairs of nodes[75].

Real-world graphs are extracted from dynamic graph datasets included in SNAP[77] and other graph visualization research[78]. We segment each dynamic graph dataset into consecutive static snapshots by slicing it with diverse fixed temporal scales (daily, weekly, monthly, yearly). Datasets that fail the empirical size and density criteria (Table 1) are excluded, yielding 12 suitable datasets.

Graph Size (Size) Graph size, defined as the number of nodes in a graph, significantly affects visual perception and readability[15]. To ensure our experiment reflects graph comparison tasks encountered in empirical settings, we adopt the categorization established in a previous survey[16]. Accordingly, graph size is categorized into four groups: small ($[10, 20]$), medium ($[21, 50]$), large($[51, 200]$), and very large ($[201, 400]$). Although the referenced survey[16] does not specify an upper bound for the very large category, we impose a limit of 400 nodes to prevent excessive variance in visual complexity within the group. To mitigate the influence of outliers, graphs within each category are sampled using Gaussian distributions centered on the category median.

Edge Density (Density) Edge density, representing the frequency of connections between nodes, also significantly influences visual perception and readability[15]. We adhere to the notion of linear density ($|E|/|V|$) and categorization of Yoghourdjian et al.[16], excluding tree-like structures that often result in disconnected components (density range $(0, 1)$). Consequently, the density categories are defined as sparse ($[1, 2)$), dense ($[2, 3)$), and very dense ($[3, 10]$). The upper density limit is capped at 10, approximating the maximum density observed in our real-world dataset (9.49). Same as the graph size, graphs within each category are sampled using Gaussian distributions centered on the category median.

Visualization Layout (Layout) As we select the node-link diagram as a visual representation of graphs, the layout of the node-link diagram has a great effect on the perception of graphs[17,55]. We employ three different graph layout algorithms in different categories based on their popularity and ease of use[35](shown in Figure 3):

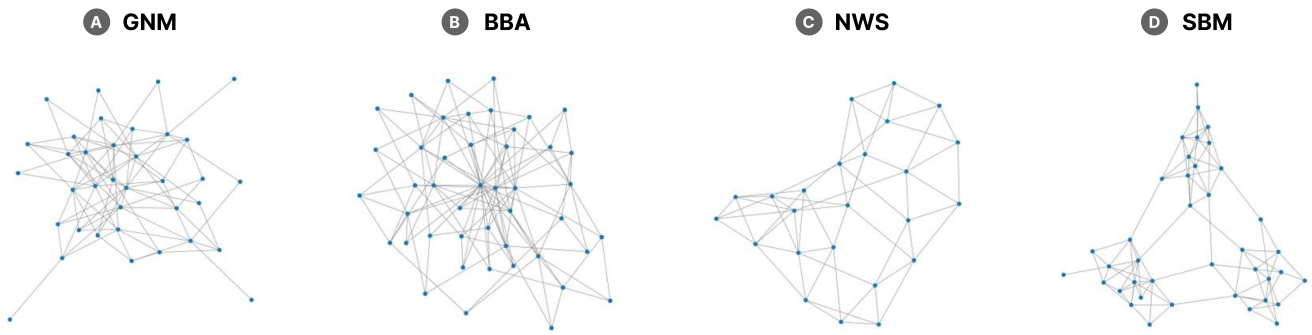

Figure 2. In this study, four different synthetic graph generation algorithms are utilized for stimuli generation. A) **GNM** algorithm randomly generates M edges between N nodes. B) **BBA** algorithm has power-law degree distributions of node degrees by connecting edges between new nodes to existing nodes with high degree [73]. C) **NWS** algorithm creates a ring over nodes and connections between their k nearest neighbors [74]. D) **SBM** algorithm partitions a graph into blocks of arbitrary sizes whose edges are placed between pairs of nodes [75].

- **Force-directed Layout (Fruchterman-Reingold)**: An algorithm that simulates a physical system where nodes act as repelling particles and edges act as attracting springs. This algorithm creates an aesthetically pleasing layout with uniform edge lengths, effectively revealing symmetries and clusters [79].
- **Circular layout**: This layout positions all nodes equidistantly along the circumference of a circle. It provides a structured view that highlights edge density and connectivity patterns across the graph, without the node occlusion [80].
- **Multidimensional scaling layout (UMAP)**: Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction technique. When applied to graphs, it projects the topological structure into a 2D space, effectively preserving both local neighborhood and the global structure [81].

Visual Encodings and Standardization To mitigate confounding factors from low-level visual variations, we standardize visual encodings using a systematic scaling strategy. Node sizes and edge widths are proportionally reduced for larger or denser graphs to prevent occlusion, with identical parameters applied uniformly across all layout algorithms to ensure fair comparison. Regarding color encodings, we prioritize legibility and familiarity to ensure that nodes remain visually distinct even amidst edge clutter. To achieve this, we adopt the *Tableau10* palette, a color scheme widely used in the information visualization community, assigning the blue hue to nodes and the gray hue to edges.

Summary The final dataset included 1,881 node-link diagrams: 1,152 synthetic and 729 real-world graphs. Although some conditions fall short of the target of 32 graphs due to their inherent sparsity and multi-component structures (particularly very large and very dense cases), sufficient diversity in visual comparisons is maintained. Specifically, except for the very large-very dense condition, which yielded no graphs, at least six graphs were collected for all other conditions, allowing for a sufficiently diverse set of comparisons. Dataset can be found online[1].

Experiment 1: Assessing Human Perception

This section describes Experiment 1, including the collection of human subject data from visual graph comparison tasks and the analysis of the results. The primary objective is to answer the first research question: *Can humans discern similarity and differences between graph visualizations, and which factors significantly affect these decisions?*

To answer this question, we first examine whether a consensus exists in human graph similarity assessments that can be evidence that humans can discern the similarity between graphs. Furthermore, we investigate whether human perception extends beyond simple binary differentiation to capturing the varying degrees of similarity. Finally, we identify the specific visual features underpinning these cognitive processes and evaluate how perceptual performance varies across three graph attributes: graph size, density, and layout.

Experiment Design

Ground Truth Establishment Ground truth is essential for measuring accuracy and evaluating human ability. However, defining an objective ground truth for graph similarity is challenging due to the lack of a universal mathematical definition or verified computational proxies [82]. In this circumstance, when algorithmic metrics are insufficient, we use human interpretation as a proxy for ground truth [83]. We establish ground truth using a visual clustering strategy, assuming that intra-cluster similarity exceeds inter-cluster similarity [84]. For synthetic graphs, we rely on generation algorithms to distinguish visual clusters because they yield distinct topological patterns (see Figure 2). Real-world graphs are manually clustered by three authors, including the first author. To avoid bias, the authors perform this task based solely on holistic visual resemblance without any predefined criteria or guidelines. The individual clustering results are aggregated, and the final groups are established by majority consensus, ensuring that only graphs consistently perceived as similar are grouped together.

Task Design The experiment employs a triplet, a query graph, and two target graphs for the relative graph comparison task. When given a triplet, participants are asked to select the target most visually similar to the query, without any available interactions such as zooming or panning. This restriction is given to maintain strict experimental control by ensuring consistent viewing conditions across all participants, while also prioritizing holistic perceptual

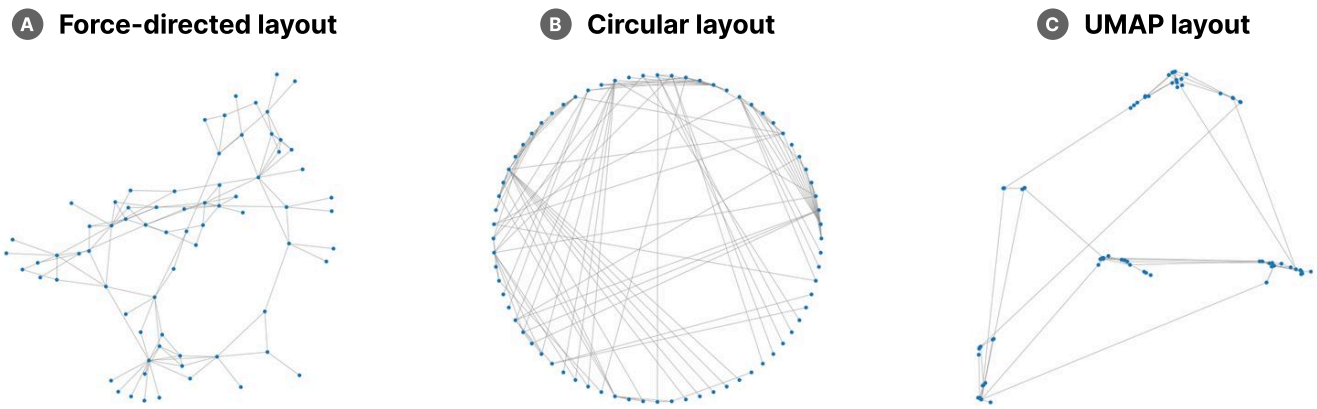

Figure 3. Node-link diagrams of a real-world graph drawn with three different graph layout algorithms utilized in this study. A) **Force-directed layout (Fruchterman-Reingold)** creates an aesthetically pleasing layout with uniform edge lengths by simulating a physical system where nodes act as repelling charged particles and edges act as attracting springs [79]. B) **Circular layout** positions all nodes equidistantly along the circumference of a circle, providing a structured view that highlights edge density and connectivity patterns across the graph [80]. C) **Multidimensional scaling layout (UMAP)** projects the topological structure into a 2D space with the UMAP algorithm, effectively preserving both local neighborhood relationships and the global structural organization [81].

judgments over overly detailed analysis. Then, participants are asked to justify their choice based on six predefined criteria and rate their confidence on a 5-point Likert scale.

This design moves beyond traditional binary choices to capture nuanced similarity decisions [36], adopting a relative comparison task from a previous study on the similarity perception of timeseries visualizations [37]. Confidence ratings allow quantification of perceived similarity, from indistinct differences (1) to clear visual discriminability (5).

Regarding visual arrangement, we place the query graph at the center and juxtapose the target graphs side by side. Under the constraints of this study, this choice is the most appropriate for performing the task among other alternatives: superposition and explicit encoding [14]. Superposition, one of the design alternatives, is infeasible because our dataset comprises graphs without node correspondence, making alignment computationally prohibitive (NP-hard) [85]. Likewise, we exclude explicit encoding to maintain the study's scope on baseline node-link diagrams without auxiliary visual encodings.

To minimize experimental noise and avoid the decision paralysis observed in pilot studies where both target graphs originated from clusters distinct from the query, we strictly control the questionnaire generation process to ensure that each triplet contains at least one target graph from the same visual cluster as the query graph, thereby guaranteeing a baseline level of visual similarity. Therefore, the triplets are categorized into two conditions: the *Same-Group condition*, where all three graphs are drawn from the same cluster, and the *Distinct-Group condition*, where exactly one target originates from a different group. Finally, to ensure randomness and eliminate positional bias, the on-screen placement of the targets is randomized for each trial.

Aligning Graph Orientation Furthermore, the questionnaires are designed to minimize errors arising from the stochastic nature of graph layout algorithms and the inherent characteristics of human visual perception. Layout algorithms, particularly force-directed models, can produce varying outputs for the identical graph structure [79]. Even if the same graph data is presented in an identical layout, variations in orientation alone can cause participants to perceive them as distinct graphs [86].

Previous research preserves the user's mental map by minimizing node displacement between pairs, which is crucial for accurate similarity assessments [19,70]. However, direct node alignment is infeasible due to the UNC. Alternatively, we employ a computer-vision-based Intersection over Union (IoU) approach [87,88]. We align graph centroids and apply dilation to calculate the IoU Area Under the Curve (AUC). This metric quantifies the pixel-wise similarity between the graphs. To find the alignment that yields the highest similarity, we iteratively rotate the target graphs by 10° increments to identify the orientation that maximizes the AUC relative to the query graph.

Criteria of Human Decisions Humans employ various visual criteria when assessing graph similarity. To identify the primary criteria used in these decisions, previous studies either encourage open-ended responses with qualitative analysis [89] or provide predefined criteria to facilitate structured responses [90]. We adopt the latter method, providing participants with clearly defined criteria derived from existing literature. This structured approach helps participants articulate their reasoning consistently and precisely. The finalized criteria presented to participants are as follows:

- *Overall Shape*: Global visual arrangement or silhouette of the graph.
- *Local Shapes*: Specific structural patterns or motifs formed by subsets of nodes and edges.
- *Graph Size*: Total number of nodes.
- *Node Degrees*: Distribution and prominence of highly connected (high-degree) nodes.
- *Edge Density*: The Extent to which edges are densely or sparsely connected among nodes.
- *Communities*: Identifiable clusters or groups of nodes forming distinct substructures.

To develop these criteria, we review previous research investigating visual similarity perception in graphs. Ballweg et al. [91] found that participants frequently relied on overall

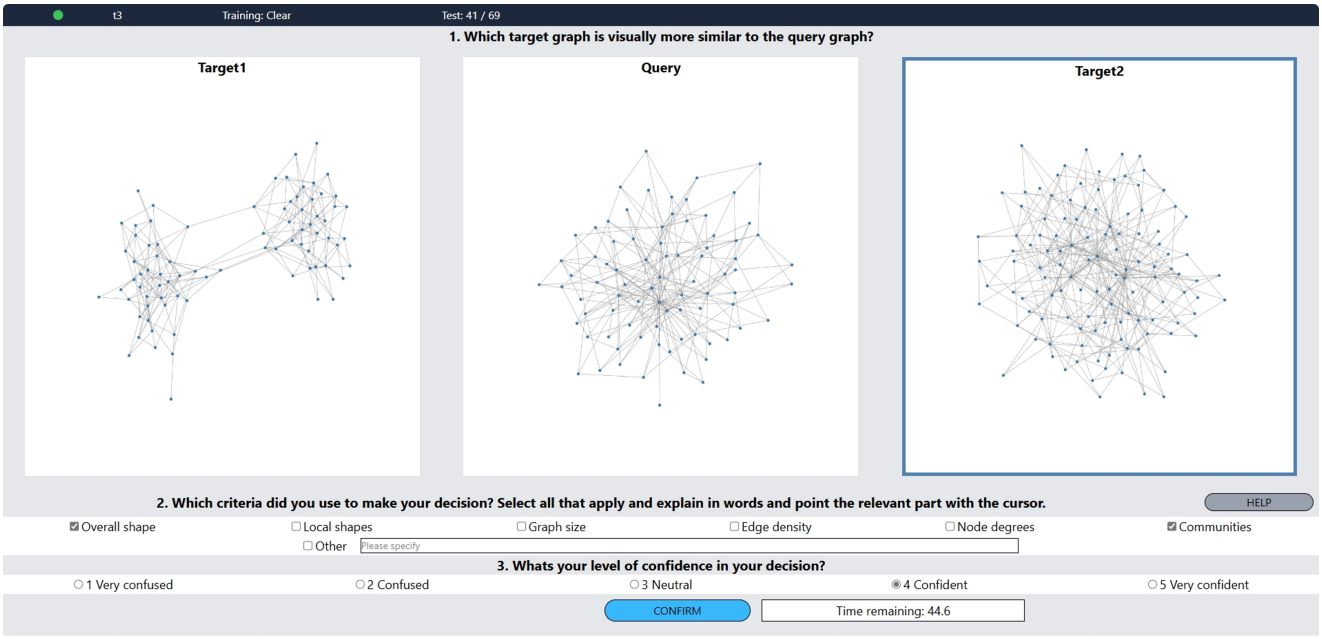

Figure 4. The system employed in Experiment 1 is designed to collect human graph similarity assessment data. For each question, three node-link diagrams are presented. Participants answer the questions in the following order.1) The user selects the target graph that seems more similar to the central query graph. 2) Next, they choose and explain the decision criteria from the options below. 3) They then indicate their confidence in their choice. 4) The entire process must be completed within one minute. If additional clarification on the criteria is needed, the user can press the Help button to review the explanation.

graph shapes, hierarchical structures, and node distribution patterns when judging Directed Acyclic Graphs. Bridgeman and Tamassia[92] demonstrated significant influences from global layout consistency, graph size variations, and edge structural changes on perceived graph similarity. Von Landesberger et al.[56] further supported these findings, noting that local structural variations such as appearance or disappearance of specific connections, edge density, node centrality, and clearly defined community structures significantly influenced similarity decisions.

Experiment Process

This study is approved by the Institutional Review Board of our institution (IRB No. 2501/004-009). 32 adult participants (26 males, 6 females), aged 22 to 35 (mean age 26.6), are recruited. Thirty participants hold university degrees (BSc, MSc, PhD), all reporting intermediate familiarity with graph data and visualizations. Experiments are conducted in-lab or via Zoom using standard desktop or laptop computers with a mouse. Each session lasts approximately one hour, including introductions, training, and tasks. Participants provide informed consent and receive instructions detailing the research procedure. Instructions include an introduction to the tasks participants are required to complete in the three stages, along with explanations and example figures illustrating the criteria to be selected during the second stage.

After the instruction, participants complete three training tasks using additional unused data before the main experiment. In the actual experiment, each participant performs two tasks for each source type (real-world and synthetic) across the 36 combinations of independent variables. However, due to the exclusion of the three very large and very dense real-world cases, participants complete a total of 69 tasks individually. Task order is counterbalanced using a Latin square design.

Participants are encouraged to make decisions within 60 seconds per task to avoid overly detailed analysis[93]. They are explicitly informed that no guidance is provided on similarity decisions or classification criteria, ensuring the responses genuinely reflect participants' abilities to interpret graphs visually. Instead, participants can revisit the instruction describing the criteria via a help button (see Figure 4), and explanations for each criterion are provided immediately.

Results and Findings

Consistency in Human Decision To assess whether humans can perceive visual similarities in graphs consistent with the ground truth, we conduct a one-sample t-test of participants' accuracy against the chance level of 0.5 (Figure 5). The analysis reveals that participants achieve a mean accuracy of 77.2% ($M = 0.772, SD = 0.086$), which is significantly higher than the random chance level ($0.5, p < .001$). Notably, 28 out of 32 people show accuracy significantly above the chance level, and even the participant with the lowest performance (56.1%) performs above chance. Moreover, the relatively low standard deviation indicates a high level of consistency in visual graph perception across the population. These results strongly support the hypothesis that humans share a consensus on visually distinguishing graph similarities.

To determine which independent variables influence these human decisions, we perform a three-way ANOVA examining the effects of graph size, edge density, and layout on accuracy. The results show that graph size is the only factor exerting a statistically significant main effect on accuracy ($p = .046$). However, post-hoc analysis using Tukey's HSD fails to identify significant differences

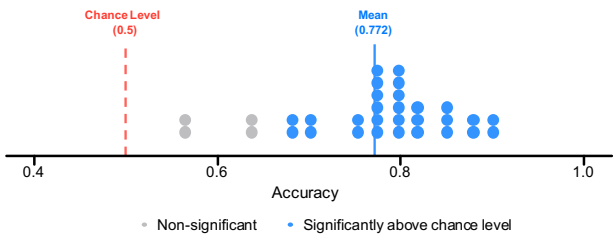

Figure 5. Accuracy distribution of participants in Experiment 1. The red dashed line represents the random chance level (0.5). Each dot represents an individual participant. Based on a one-sample t-test ($H_0 = 0.5$) for each participant, blue dots denote those who exhibited accuracy levels significantly above chance (28 out of 32, $p < .05$), while gray dots represent those whose performance was statistically indistinguishable from chance. These results demonstrate a robust human capacity to visually distinguish graph similarities.

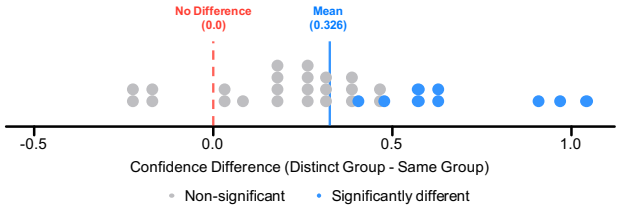

Figure 6. Distribution of confidence differences between Distinct-Group and Same-Group conditions. For each participant, confidence scores (rated on a 1–5 scale) are averaged within each condition, and the mean score of the Same-Group is subtracted from that of the Distinct-Group. The red dashed line marks the baseline of no difference (0.0), and the blue solid line indicates the overall mean difference (0.326). Blue dots represent participants who show a statistically significant difference between two conditions (Mann-Whitney U test, $p < .05$), while gray dots indicate no significant difference. Although the global mean shows a positive trend, the majority of participants (23 out of 32) do not reach statistical significance individually, highlighting substantial heterogeneity in sensitivity to graph differences.

between any specific pairs of size groups. This discrepancy can be attributed to the borderline significance of the ANOVA result and the conservative nature of Tukey's HSD test, which imposes stricter thresholds for pairwise distinctions, indicating that the effect of size is relatively subtle and not driven by distinct disparities across specific groups. Collectively, these findings imply that, although humans generally struggle to interpret graphs that become excessively large or dense due to factors such as the hairball effect[16], their capacity to perceive graph similarity remains robust, persisting across diverse conditions of size, density, and layout.

Consistency in Human Confidence To further investigate whether participants' visual perception scales with the magnitude of graph differences, we analyze the self-reported confidence scores, ranging from very confused (1) to very confident (5). We hypothesize that participants will report higher confidence in trials in which the target graphs belong to distinct groups (Distinct-Group condition) compared to trials where both targets are from the same group as the query (Same-Group condition). This hypothesis is grounded in the premise that the Distinct-Group condition exhibits more salient visual differences compared to the Same-Group condition, thereby enabling participants to select the target graph from the same group with greater confidence.

To validate this, we verify whether the confidence levels in the Distinct-Group condition are statistically higher than those in the Same-Group condition using a Mann-Whitney U test (Figure 6). The global analysis reveals a statistically significant difference between the two conditions ($U = 665,328.0, p < .001$). When calculated on a per-participant basis, participants exhibit higher confidence in the Distinct-Group condition ($M = 3.548, SD = 0.397$) than in the Same-Group condition ($M = 3.221, SD = 0.376$), with a mean difference of 0.326.

However, a more granular analysis performed at the individual level reveals considerable variability. When independent Mann-Whitney U tests are conducted for each participant, only 9 out of 32 participants (28.1%) demonstrate a statistically significant increase in confidence for the Distinct-Group condition ($p < .05$). The remaining participants do not exhibit a significant difference, and notably, four individuals show inverse trends.

These findings suggest that while humans, at a population level, possess the perceptual sensitivity to distinguish not just the similarity but also the magnitude of differences between graphs, this capability is not uniformly distributed. The contrast between the global significance and the individual results highlights a substantial heterogeneity in individual sensitivity to visual graph differences.

Rationale of Decision Participants identify *Overall Shape* as the most dominant feature driving their similarity decisions, accounting for the highest proportion of responses (898). *Edge Density* follows as the second most frequent criterion (781), showing a notably higher count compared to other criteria. In contrast, *Graph Size* is the least selected criterion (227). These findings suggest that humans prioritize the global silhouette of the graph and readily detect patterns of edge concentration (density), whereas they are less adept at estimating the number of nodes during visual comparison. Total number of decision criteria selected by humans is shown at Table 4 with MLLMs' responses.

To investigate the influence of three independent variables on the frequency of each criterion, we perform a three-way ANOVA. The results reveal distinct interaction patterns: size significantly affects the selection frequency of all criteria except *Graph Size* itself and *Communities*. Layout significantly influences all criteria except *Local Shapes* and *Graph Size*. Density, in contrast, exhibits a limited scope of influence, significantly affecting only the selection of *Edge Density* without impacting the other criteria

Takeaways

The results confirm that humans possess the fundamental capability to visually judge graph similarity, supporting the feasibility of visual comparison tasks. However, while collective human perception reliably reflects the ground-truth difficulty, individual sensitivity to the strength of graph similarity is heterogeneous. Some users are highly

sensitive to the magnitude of differences, while others may perform the comparison task accurately but with a flatter confidence profile regardless of task difficulty. Therefore, we recommend designing comparison interfaces that rely on relative judgments rather than requiring absolute quantification of similarity.

Regarding the decision rationale, the overwhelming reliance on *Overall Shape* implies that preserving the global topology is paramount in comparative visualization. Systems should prioritize rendering techniques that emphasize the global silhouette. Furthermore, *Edge Density* emerges as a critical factor, suggesting that visual patterns created by edge concentrations and resulting occlusions serve as key discriminatory cues.

Among the independent variables, size has the greatest influence on perception, followed by layout and density. Interestingly, accuracy tends to be lower for small graphs. We hypothesize that small graphs lack sufficient node volume to form a distinct global silhouette, making *Overall Shape*, the primary human decision criterion, less reliable. In such cases, alternative visual aids may be necessary. Conversely, for medium to very large graphs, the increased number of nodes facilitates the formation of unique global structures, allowing users to effectively distinguish graphs based solely on their *Overall Shape*.

Experiment 2: Deriving Human-Measure Relationships

In this experiment, we aim to answer the second research question: *Do traditional computational measures capture the similarities and differences that humans perceive visually?* To address this, we evaluate the alignment between various computational metrics and human decisions. Our objective is to identify measures that not only effectively approximate human perception but also facilitate interpretable comparisons within visual analytics systems. Furthermore, such measures can serve as reliable Visual Quality Measures (VQMs) for graph similarity[69].

Experiment Design

The selection of computational measures is constrained by the target graph specifications defined in our study design. Specifically, the measures must support comparisons between undirected, unweighted graphs without node correspondence (UNC). Under these constraints, we select a total of 16 graph similarity measures. A list of the selected measures is provided in Table 2.

First, we include four graph-attribute-based measures that correspond directly to the quantifiable visual criteria identified in Experiment 1. Among the six criteria used by humans, *Overall Shape* and *Local Shapes* are excluded due to the difficulty of direct quantification. For the remaining four attributes (Graph Size, Edge Density, Node Degrees, and Communities), we derive similarity scores using two metrics. *Balance*, defined as the ratio of the minimum to the maximum value, is utilized to calculate the similarity of size and density. *Divergence*, calculated as the Jaccard similarity between distributions, is for comparing the distribution of node degrees and sizes of communities detected by the Louvain algorithm.

Table 2. List of graph similarity measures employed in Experiment 2. Size and density balances are computed as the ratio of the maximum value to the minimum value. Node degree and community divergences are calculated by computing the Jaccard similarity between the node degrees and the sizes of the communities detected by the Louvain community detection algorithm. The remaining measures are implemented according to the descriptions provided in the references or by using available code. All measures are normalized to yield values between 0 and 1, with higher numbers indicating greater similarity between graphs.

Category	Measure
Attributes	Size balance, Node degree divergence, Density balance, Community divergence
Embedding	Netsimile[94], Portrait divergence[95]
Spectral	Laplacian spectral[44], Feather[96], Ipsen-Mikhailov[99], NetLSD[72]
Graphlets	GCD-11[46], Netdis[100]
Alignment	REGAL[101], GRASP[102]
Kernels	Shortest-path kernel[103], Weisfeiler-lehman kernel[104]

The remaining 12 measures are selected based on our *literature review* to ensure a diverse representation of theoretical foundations. Each measure is characterized by distinct properties, including embedding-based[94,95], graphlet-based[46,71], spectral-based[44,72,96], graph kernel-based[45,97], and graph alignment-based approaches[98].

To ensure a consistent representation of similarity across all measures, measures that originally denote distances are normalized to a similarity scale in which 0 denotes complete dissimilarity and 1 denotes identity. Specifically, distinct transformation rules are applied based on the theoretical range of each metric: for measures strictly bounded within $[0, 1]$, the transformed similarity is calculated as $1 - d$. For measures defined on the unbounded range $[0, \infty)$, the transformation $1/(1 + d)$ is employed.

Experiment Process

Unlike human participants who perform relative comparisons due to the inherent difficulty of absolute quantification, computational measures provide absolute similarity scores for any given pair of graphs. For each triplet consisting of a query graph (Q) and two target graphs (T_A, T_B), we calculate the computational similarity scores $S(Q, T_A)$ and $S(Q, T_B)$. We then assess whether the choices made by human participants align with the higher similarity score. This process enables us to quantify the concordance rate between algorithms and human visual perception.

Results and Findings

Agreement with Human Judgment To identify the computational measure that best approximates human graph-similarity perception and thus enables interpretable graph comparisons, we analyze the agreement between human judgments and the 16 selected metrics. We employ Cohen's Kappa (κ) as the primary evaluation metric, as it accounts for the possibility of random agreement, thereby providing

a more robust estimate of concordance than accuracy[105]. Agreements of all 16 measures are listed at Table 3.

The analysis reveals that **Portrait divergence**[95] demonstrates the best performance. It achieves a κ value exceeding 0.4, which is interpreted as a moderate level of agreement according to established benchmarks[105]. Several other measures also achieve κ values near or above 0.4, falling within the margin of error of **Portrait divergence** (**Ipsen-Mikhailov**[99], **NetDis**[100], and **NetSimile**[94]). In contrast, the remaining measures generally produce values below 0.3, with **GRASP** recording the lowest ($\kappa = 0.1477$), suggesting a negligible agreement with human perception.

To further validate the reliability of the top-performing measures, we conduct an ANOVA to investigate how the three independent variables, graph size, edge density, and layout, affect their agreement with human decisions. The results indicate that **Portrait divergence** is not significantly influenced by any of the three variables, demonstrating robust agreement across all experimental conditions. Conversely, other top contenders are affected by specific factors: **Ipsen-Mikhailov** by layout, **NetDis** by graph size, and **NetSimile** by both size and layout.

Additionally, we examine the relative rankings for each participant to account for the inherent heterogeneity of human perception. **Portrait divergence** achieves the highest stability with an average rank of 4.19, placing as the best for 9 participants and within the top four for the other 11 participants. Wilcoxon Signed-Rank tests confirm its significant superiority ($p < .05$) over most measures, except for the three other top contenders, verifying its consistency across diverse observers.

In conclusion, **Portrait divergence** proves to be the most reliable measure, exhibiting superior stability in terms of both the strength of agreement and robustness against variations in graph properties and individual differences.

Correlation with Human Confidence In this analysis, we compute Spearman's correlation (ρ) between human confidence ratings and the magnitude of the difference in computational similarity scores. The objective is to verify whether the numerical gap between computational scores reflects the degree of certainty perceived by humans. Correlations of all 16 measures are listed at Table 3.

A direct comparison presents a methodological challenge due to the disparate nature of the data: human confidence is recorded on a discrete, positive 5-point Likert scale, whereas computational measures yield continuous difference values that can be negative. To reconcile these disparate scales, we apply a systematic transformation process. We first compute the absolute difference between the similarity scores of the two target graphs ($|S(Q,T_A) - S(Q,T_B)|$) to eliminate sign ambiguity. Subsequently, these absolute values are normalized to the $[0,1]$ range using Min-Max scaling and discretized into five equal-width bins (0.2 intervals) to align directly with the 5-point human confidence scale.

Consistent with the agreement analysis, **Portrait divergence** exhibits the highest correlation with human confidence, achieving a Spearman coefficient (ρ) of 0.2677. Although this value falls slightly below the conventional threshold of 0.3 typically required to indicate a weak correlation[106], it is statistically significantly higher than that of all other competing measures.

Table 3. Agreement (Cohen's κ) of 16 measures with human perception in selecting the target graph more similar to the query, and the Spearman's correlation (ρ) between human confidence and the absolute difference of measured similarities $|S(Q,T_A) - S(Q,T_B)|$. Measures are ranked in descending order of agreement (κ). **Portrait divergence** achieved the highest performance in both metrics (bold). While three other measures (**Ipsen-Mikhailov**, **Netdis**, **Netsimile**) record agreement scores comparable to **Portrait divergence**, no other measure achieved a correlation score within a similarly competitive range.

Measure	Agree. (κ)	Corr. (ρ)
Portrait divergence	**0.4247**	**0.2685**
Ipsen-Mikhailov	0.4111	0.1967
Netdis	0.4012	0.2046
Netsimile	0.3966	0.0796
Feather	0.3576	0.1421
Shortest-path kernel	0.3352	0.0903
Node degree divergence	0.3196	0.1704
REGAL	0.3151	0.1499
GCD-11	0.3134	0.1465
Laplacian spectral	0.2862	0.1252
Density balance	0.2727	0.1093
Community divergence	0.2401	0.1861
Size balance	0.2138	0.0906
NetLSD	0.1749	0.0182
Weisfeiler-lehman kernel	0.1732	0.0472
GRASP	0.1477	0.1009

To further verify the consistency of this correlation across human participants, we analyze each participant's relative rankings of the measures. **Portrait divergence** demonstrates superior stability, ranking first for 11 participants and within the top four for 12 other participants, with an average rank of 4.34. A Wilcoxon Signed-Rank test confirms that it significantly outperforms most other measures, except **NetDis** and **Community divergence**.

However, a notable distinction emerges regarding robustness. While the agreement rate of **Portrait divergence** remains stable across all experimental conditions, its correlation with confidence is sensitive to data characteristics. ANOVA results for each measure indicate that the correlation strength is significantly influenced by graph size and edge density. Nevertheless, despite these dependencies, **Portrait divergence** remains the most effective computational proxy available for approximating human decision confidence.

Takeaways

The extent to which computational measures align with human visual perception of graph similarity varies significantly across metrics. **Portrait divergence**, identified as the top performer in both agreement and correlation, incorporates topological characteristics across all structural scales and is applicable to all network types[95]. **Ipsen-Mikhailov**, which also achieved high agreement, measures global spectral distance for comparing the overall topological structures of two networks[99]. This finding suggests that measures

reflecting global structures align best with human perception, making them the most effective measures to guide humans' visual graph comparison tasks **(RQ2)**. This is also consistent with our observation in Experiment 1 that *Overall Shape* serves as the overwhelming decision criterion for humans.

However, even the best-performing measures demonstrate only moderate capabilities, falling short of serving as ideal VQMs for graph similarity. While the agreement between human decisions and computational measures is moderate, the correlation between measure differences and user confidence remains weak. This discrepancy may stem from the consistently high confidence observed in Experiment 1, which contrasts with the variable magnitudes of computational differences. In summary, while humans can reliably determine which graph is more similar, they appear less adept at quantifying the exact degree of similarity. Consequently, there exists a clear need for alternative approaches that better bridge this gap.

Experiment 3: Assessing the Potential of MLLMs

In this final experiment, we address the third research question: *Do MLLMs possess the capability to align with human perception and guide users in graph comparison tasks?* To answer this, we employ three state-of-the-art general-purpose MLLMs to perform the graph comparison task. We assess their alignment with human perception and investigate their potential advantages over the best-performing computational measure (**Portrait divergence**) identified in Experiment 2.

Experiment Design

Given that MLLMs exhibit distinct strengths across specialized areas and benchmarks[29], it is crucial to empirically evaluate their performance specifically within the context of graph similarity assessment. To this end, our evaluation framework mirrors the design of Experiment 1, assessing the visual cognitive capabilities of MLLMs through a relative comparison task. Consistent with the methodology in Experiment 2, we measure agreement and correlation with human judgments to determine whether MLLMs can serve as proxies for human perception and provide effective guidance for users in graph comparison tasks. Furthermore, we analyze the alignment of decision criteria to gain insight into the interpretability and rationale underlying the MLLMs' decisions.

Model Selection The landscape of MLLMs is highly competitive, with rapid advancements from major providers. As of October 2025, Anthropic, OpenAI, and Google are recognized as the dominant market leaders[107]. Accordingly, we selected the latest flagship general-purpose MLLM from each of these three providers for our evaluation:

- **Google Gemini 2.5 Pro** (hereinafter **Gemini**) [2]
- **OpenAI GPT-5** (hereinafter **GPT**)[3]
- **Anthropic Claude Sonnet 4.5** (hereinafter **Claude**)[4]

Prompting To enable MLLMs to perform visual graph comparison effectively, we develop a structured prompting strategy designed to mimic the human task in Experiment 1. We assign each model the persona of an "Expert researcher in network visualization and graph drawing" and instructed it to identify which target graph (T_1 or T_2) is visually more similar to the Query Graph (Q).

To encourage reasoning and improve accuracy, the prompt required the models to follow a four-step Chain-of-Thought (CoT) process:

1. **Internal Evaluation:** Internally compare Q, T_1, and T_2 across six visual features (*Overall Shape*, *Local Shapes*, *Graph Size*, *Node Degrees*, *Edge Density*, *Communities*) to determine the winner.
2. **Rationale Formulation:** Formulate a concise explanation justifying the selection, directly stating the key differences that led to the decision.
3. **Confidence Assessment:** Assign a confidence score on a scale ranging from very confused (1) to very confident (5).
4. **Decision Criteria Contribution Array:** Provide an array of length six corresponding to the visual features. Each value in the vector should be $\{-1, 0, 1\}$, indicating whether the chosen target was inferior, equal, or superior to the alternative target regarding that specific feature.

To ensure reproducibility and deterministic outputs, the temperature parameter is set to 0 for Gemini and Claude. For GPT, the default setting is used because precise temperature control is not available.

Input Data and Preprocessing Along with the prompt, the MLLMs are provided with the three graph visualization images used in the human trials. To maintain methodological consistency with Experiment 1, we employ the identical protocols for image generation and graph alignment. Furthermore, to ensure consistent input quality and token efficiency, the images undergo standardized preprocessing. Each image is cropped to remove excess whitespace and subsequently padded to achieve a square aspect ratio.

We harmonize image resolution to meet the API specifications for each model. While **Gemini** and **Claude** support higher resolutions (768×768 and 1092×1092, respectively), **GPT** is optimized for 512×512 pixels. To ensure a fair comparison under identical conditions, all images are resized to a maximum of 512×512 pixels.

Output Format The models were instructed to generate a structured JSON output containing the following fields:

- `Decision`: The label of the chosen graph ($T1$ or $T2$) which seems more similar to the query graph Q.
- `Rationale`: A text string explaining the reason for the decision.
- `Confidence`: An integer score (1–5) implying the confidence of the decision.
- `Criteria`: An array of 6 integers representing contributions of six predefined decision criteria.

In addition to the content of the response, we record the inference latency (time elapsed from request to response) to evaluate the practical efficiency of each model.

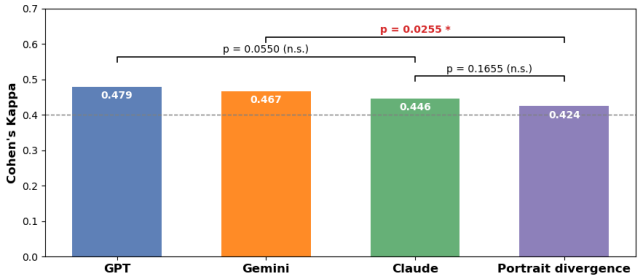

Figure 7. Bar chart comparing the agreement levels (Cohen's Kappa) of three state-of-the-art MLLMs against **Portrait divergence**. The grey dotted line indicates the moderate level of agreement ($\kappa = 0.4$). Although no statistically significant differences are observed among the MLLMs, **GPT** achieves the highest nominal performance. Notably, both **GPT** and **Gemini** demonstrate significantly higher agreement with human judgments compared to **Portrait divergence**, indicating that these models effectively surpass the capabilities of traditional computational metrics.

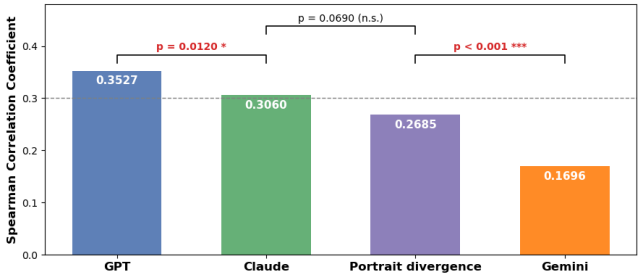

Figure 8. Bar chart comparing the Spearman's correlation (ρ) between human confidence and the decision confidence of three MLLMs, benchmarked against **Portrait divergence**. The grey dotted line indicates a weak correlation level ($\rho = 0.3$). **GPT** demonstrates a statistically significant improvement in correlation compared to **Portrait divergence**. **Claude** shows a comparable correlation with no significant difference observed. In contrast, **Gemini** exhibits a significantly lower correlation than the computational baseline, indicating a divergence from human uncertainty patterns.

Results and Findings

Agreement with Human Judgment To evaluate the alignment between MLLM decisions and human judgments, we calculate Cohen's Kappa (κ). Furthermore, to determine whether MLLMs offer a statistically significant improvement over traditional methods, we compare their performance against *Portrait divergence*, which is identified as the top-performing computational measure in Experiment 2. We employ a bootstrap method (resampling $N = 2,000$) to test the significance of the difference in Kappa coefficients. Results are shown in Figure 7.

GPT demonstrates the highest alignment with human judgment ($\kappa \approx 0.479$), showing a statistically significant improvement over **Portrait divergence** ($\kappa \approx 0.424$, $p = 0.0085$). **Gemini** also exhibits a high level of agreement ($\kappa \approx 0.467$), significantly outperforming **Portrait divergence** ($p = 0.030$). In contrast, while **Claude** achieves a respectable agreement score ($\kappa \approx 0.446$), the difference compared to **Portrait divergence** is not statistically significant ($p = 0.168$). Additionally, pairwise comparisons among the models reveal a significant performance gap between **Claude** and the other two models, whereas the difference between GPT and Gemini was not statistically significant.

Crucially, this pattern of agreement demonstrates robustness across experimental conditions. To investigate whether the superiority of MLLMs persisted under specific graph attributes, we conduct a stratified bootstrap analysis (resampling $N = 2,000$) for each independent variable: graph size, edge density, and layout. Our analysis indicates that while the relative ranking of the methods remained consistent across all variations, the performance gaps between models within individual conditions are not statistically significant. Notably, even the difference between the top-performing model (**GPT**) and the baseline (**Portrait divergence**) does not reach statistical significance when analyzed within these subdivided categories. Consequently, we interpret the statistically significant superiority observed in the aggregate analysis as the result of marginal performance gains that accumulate consistently across all conditions, rather than being driven by drastic disparities in specific scenarios.

Correlation with Human Confidence We employ Spearman's correlation coefficient (ρ) to assess whether the MLLMs' confidence scores align with the nuances of human certainty. The results, including bootstrap significance tests against **Portrait divergence**, are presented in Figure 8.

GPT exhibits the strongest correlation with human confidence ($\rho \approx 0.353$), significantly outperforming **Portrait divergence** ($\rho \approx 0.269$, $p = 0.0004$). This suggests that GPT-5 not only identifies the similar graph correctly but also mimics the human distribution of certainty.**Claude** shows a correlation ($\rho \approx 0.306$) comparable to **Portrait divergence**, with no statistically significant difference found ($p = 0.136$).Conversely, **Gemini** records a significantly lower correlation ($\rho \approx 0.170$) compared to both **Portrait divergence** ($p = 0.0004$) and GPT-5 ($p < 0.001$).

We hypothesize that **Gemini**'s lower correlation may be attributed to the temperature setting. To ensure reproducibility, **Gemini** is queried with a temperature of 0, which tends to force the model into making highly confident, deterministic decisions, thereby reducing the granularity of its confidence scores compared to humans. In contrast, **GPT**, where precise temperature control is unavailable, likely operates with a default non-zero temperature, resulting in a more natural distribution of confidence scores that better mirrors human uncertainty.

Interestingly, **Claude** maintains a relatively high correlation ($\rho \approx 0.306$) despite also being queried at a temperature of 0. This contrasts with **Gemini** suggests that **Claude** possesses superior internal calibration, allowing it to express nuanced uncertainty even under deterministic decoding strategies. Unlike **Gemini**, which is prone to overconfidence bias in this setting, **Claude**'s alignment appears to better preserve the probabilistic granularity akin to humans.

Similar to the agreement analysis, these correlation trends are robust across all conditions of the independent variables. Consequently, the models maintain their respective performance levels regardless of the visual or structural complexity of the target graphs.

Rationale of Decision To evaluate whether MLLMs' reasoning processes align with human perception, we

Table 4. Criteria selected as decision rationale by humans and MLLMs. While humans tend to select only the most salient features, resulting in a lower total count, MLLMs report a much broader and more frequent range of criteria. The models show strong support for human judgments regarding *Overall Shape*, *Edge Density*, and *Node Degrees*; however, considerable conflicts are observed in *Local Shapes*, *Communities*, and most notably, *Graph Size*.

Criteria (Human)	Model	Count	TP	FN
Overall Shape (898)	GPT	1835	815	83
	Gemini	1723	788	110
	Claude	2112	880	18
Local Shapes (534)	GPT	1599	390	144
	Gemini	1657	427	107
	Claude	1785	439	95
Graph Size (227)	GPT	921	124	103
	Gemini	1153	142	85
	Claude	1178	121	106
Node Degrees (541)	GPT	1950	492	49
	Gemini	1910	493	48
	Claude	1752	446	95
Edge Density (781)	GPT	2004	735	46
	Gemini	1991	725	56
	Claude	2007	741	40
Communities (466)	GPT	1261	377	89
	Gemini	1123	359	107
	Claude	981	333	133

first quantify the frequency with which MLLMs select each decision criterion, and subsequently analyze whether MLLMs identify a specific feature as a decision factor when human participants select it. The frequencies of decision criteria selected by humans and MLLMs, along with their interrelationships, are presented in Table 4.

All MLLMs select the decision criteria significantly more frequently than humans, identifying approximately 2.8 times more criteria on average (Human total: $3,447$ vs. MLLM range: $9,557 - 9,815$). This disparity likely stems from the differing nature of the reporting task: whereas human participants typically select only the primary or secondary factors influencing their decisions, models typically report every criterion that contributed, even marginally, to their output.

However, distinct variations are observed in the frequency of selection across different criteria. All three models report relatively low frequencies for *Graph Size* ($921 - 1,178$) and *Communities* ($981 - 1,261$). Conversely, *Overall Shape* ($1,723 - 2,112$), *Local Shapes* ($1,599 - 1,785$), *Node Degrees* ($1,752 - 1,950$), and *Edge Density* ($1,991 - 2,007$) are selected significantly more often. This suggests that, similar to humans, MLLMs regard *Overall Shape* and *Edge Density* as the most salient factors in visual graph comparison.

Furthermore, the models demonstrate a heightened sensitivity to *Local Shapes* and *Node Degrees*, features often overlooked by humans. Since these features pertain to specific substructures or local details rather than the overall gestalt, this observation is consistent with findings from related studies suggesting that MLLMs exhibit a more pronounced local focus, whereas human observers emphasize global patterns [108].

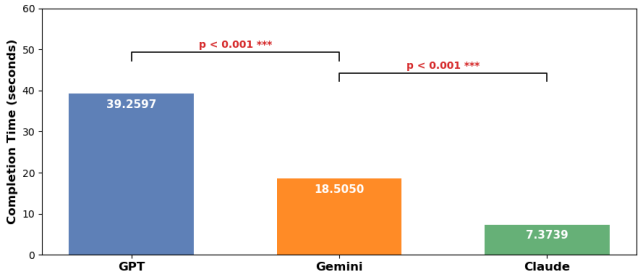

Figure 9. Bar chart depicting the mean of inference latency (completion time) for the three evaluated MLLMs. **GPT** exhibits the highest computational cost and variance (Mean: 39.26s), highlighting a trade-off between its superior reasoning capability and processing speed. In contrast, **Claude** demonstrates superior efficiency with the lowest latency (Mean: 7.37s) and a compact distribution. This positions **Claude** as the optimal choice for time-critical visual analytics scenarios.

When analyzing alignment rates, the generally higher number of criteria selected by MLLMs results in a higher frequency of both mutual agreement (True Positives) and cases in which the model selects a criterion not cited by the human (False Positives). However, conflicts arise when a model fails to identify a criterion that is salient to the human observer (False Negative), potentially causing critical confusion and undermining trust.

Regarding False Negative rates, *Overall Shape* ($2.0\% - 12.2\%$), *Node Degrees* ($8.8\% - 17.5\%$), and *Edge Density* ($5.1\% - 7.1\%$) exhibit relatively low rates. In contrast, *Local Shapes* ($17.8\% - 27.0\%$) and *Communities* ($19.1\% - 28.5\%$) show moderate rates, while *Graph Size* demonstrates a distinctively high False Negative rate ($37.4\% - 46.7\%$). This discrepancy aligns with the findings from Experiment 2, where the computational measures for **Size balance** and **Community divergence** yielded some of the lowest agreement scores. Consequently, these results reinforce the insight that humans are inherently less proficient at visually quantifying nodes or classifying communities compared to detecting topological or density-based features.

Completion Time We analyze the inference latency (completion time) for each model to assess practical feasibility. **GPT** requires the most computational time (Mean: 39.26s, Max: 142.24s), indicating a trade-off between its high accuracy and latency. **Gemini** is considerably faster (Mean: 18.50s), completing tasks in less than half the time of **GPT** on average. **Claude** demonstrates superior efficiency (Mean: 7.37s), solving tasks significantly faster than both competitors. This positions Claude as a strong candidate for scenarios where real-time processing is critical. Results are shown in Figure 9.

Takeaways

Our experiment results suggest that state-of-the-art MLLMs generally serve as effective proxies for human graph similarity perception, often surpassing the best available computational measures. Grounded on them, we derive the implications for designing visual analytics systems:

For applications that prioritize maximal alignment with human perception, **GPT** is the optimal choice. It demonstrates statistically significant superiority in both decision agreement and confidence correlation. Although it incurs a higher latency, its ability to provide high-quality rationales alongside accurate judgments makes it a competitive assistant for reducing the cognitive load in complex analysis tasks.

If the primary goal is accurate decision-making (Agreement) rather than calibrating confidence levels, **Gemini** offers a compelling alternative. It significantly outperforms traditional measures in agreement while being twice as fast as **GPT**. However, practitioners should be aware of its tendency toward overconfidence (lower correlation).

For time-sensitive or large-scale applications, **Claude** is the most suitable option. While its accuracy is comparable to that of the best computational measure (**Portrait divergence**), rather than superior, **Claude** offers the distinct advantage of providing interpretable textual explanations while maintaining extremely low latency.

Finally, all three MLLMs exhibit consistent reasoning patterns. They provide robust support for human judgments based on *Overall Shape*, *Edge Density*, and *Node Degrees*, whereas they demonstrate relatively high rates of disagreement for *Local Shapes*, *Communities*, and *Graph Size*. Therefore, when leveraging MLLM reasoning to assist human analysts, we recommend a dual strategy: models can reinforce human decisions regarding the former features, whereas for the latter, they can be utilized to highlight potential discrepancies or aspects that human observers may have overlooked.

Discussion

In this section, we propose practical guidelines for incorporating MLLMs into visual analytics systems designed for graph comparison with an example scenario. Furthermore, we outline future directions to enhance the generalizability, robustness, and interpretability of graph comparison. Ultimately, we aim to contribute to the efficiency of visual analytics by establishing a more grounded understanding of graph similarity perception.

Guidelines for Designing VA Systems for Graph Comparison Tasks with MLLMs

Based on the empirical findings from our three experiments, we propose the following guidelines for designing visual analytics systems, with a particular focus on applications involving the analysis of temporal evolution and anomaly detection in dynamic graphs.

Guideline 1. Prioritize Relative Comparison Systems should avoid tasks that require users to quantify the exact magnitude of graph changes. Experiment 1 revealed that while humans reliably identify the more similar graph (relative judgment), they struggle to consistently quantify the degree of similarity.

- **Example Scenario:** In a visual analytics system for monitoring network traffic evolution, instead of asking users to "rate the severity of the change between t_1 and t_2," the interface should present a ranked list of time steps relative to a baseline. This allows users to leverage their relative perception to identify outliers or anomalies effectively.

Guideline 2. Explicitly Encode Quantitative Attributes Do not rely solely on node-link diagrams for comparisons involving *Graph Size* or *Communities*. Our findings suggest that humans find it difficult to visually identify these attributes in graph visualizations, resulting in the lowest frequency in decision criteria and alignment with computational measures. Fortunately, since these metrics are quantifiable, they should be explicitly represented through auxiliary visualizations rather than through implicit visual encoding.

- **Example Scenario:** When visualizing the temporal evolution of a social network, rather than expecting the analyst to notice a 10% increase in node or community count, the system should pair the node-link view with a synchronized line chart or bar plot explicitly showing the number of nodes (*Graph Size*) and community counts (*Communities*) over time.

Guideline 3. Emphasize Global Structure and Density For comparing similar graphs, visualization techniques must prioritize layout stability to preserve the global silhouette[19]. As our findings in Experiment 1 indicate that *Overall Shape* and *Edge Density* are the overwhelming criteria for human judgment, preserving these features is crucial for emphasizing similarity.

- **Example Scenario:** In a dynamic graph dashboard tracking disease spread, the system should employ mental map preserving layout algorithms. This ensures that the overall shape (*Overall Shape*) and density (*Edge Density*) of the cluster remain stable across consecutive time steps, allowing analysts to instantly perceive genuine structural changes without additional cognitive load.

Guideline 4. Select Perception-Aligned Metrics for Computational Guidance To support human analysts effectively, utilize computational measures that capture global topology, such as **Portrait divergence**. Employing measures with high alignment improves the interpretability of automated recommendations, as the system's similarity aligns with what the human eye perceives.

- **Example Scenario:** When highlighting significant structural shifts across a sequence of temporal snapshots, the system utilizes **Portrait divergence** to measure similarity. This ensures that when an analyst drills down to investigate the source of a large discrepancy, the algorithmic difference corresponds to a visually salient feature, allowing the user to verify the change intuitively.

Guideline 5. Strategic Model Selection: Balancing Fidelity and Latency Practitioners must balance the trade-off between perceptual fidelity and latency. **GPT-5** is optimal for in-depth analysis requiring maximal alignment with human intuition, whereas **Claude Sonnet 4.5** is best suited for real-time monitoring or large-scale screening due to its superior speed and competitive accuracy.

- **Example Scenario:** Use **Claude** as a real-time filter to flag anomalies, then switch to **GPT-5** for a comprehensive, post-hoc forensic analysis of the identified events.

Guideline 6. Adaptive Reasoning Strategy: Reinforcement vs. Augmentation Systems should adopt a dual strategy for MLLM-generated explanations. Reasoning should reinforce user confidence on global features (*Overall Shape*, *Edge Density*) where human-model alignment is high, while augmenting analysis by highlighting local details (*Graph Size*, *Communities*, *Node Degrees*, *Local Shapes*) that humans frequently overlook.

- **Example Scenario:** The system validates the user's impression of global shape (Reinforcement) while simultaneously alerting them to local substructure deviations that might be overlooked (Augmentation).

Expanding the Scope of Graph Comparison

To improve the generalizability of our findings, future research must broaden the scope regarding graph conditions and similarity measures. In this study, we focused on the most fundamental graph format—undirected, unweighted, single-component graphs without self-loops—to establish a baseline for comparison. However, even minor topological alterations can significantly impact perceptual outcomes. For instance, introducing edge weights necessitates additional visual encodings, such as edge thickness or color. It remains an open question how strongly these salient visual cues might dominate structural perception or alter similarity judgments compared to the unweighted graphs used in this study.

Beyond diversifying graph types, there is substantial room to explore alternative similarity assessment methods. Previous surveys[40,41] indicate that measures leveraging Known Node Correspondence (KNC) tend to offer higher discriminatory power than the Unknown Node Correspondence (UNC) measures used in this work. While KNC was outside our current scope, investigating it would require a fundamental redesign of the experimental framework, such as visibly incorporating node labels to facilitate element-wise comparison. Such an extension would provide deeper insights into how semantic information interacts with topological structure during similarity assessment.

Therefore, future work should aim to replicate these experiments with broader similarity measures and complex graph types. Such expansion is essential to ensure that the proposed guidelines and MLLM proxies remain robust across a wider range of application scenarios.

Enhancing the Robustness of Quantitative Analysis

For our quantitative analysis, we generate a substantial dataset of over 1,000 graph visualizations and collected more than 2,000 human responses. While this volume is sufficient for a preliminary assessment, increasing the sample size is necessary to draw stronger statistical conclusions. Specifically, we encounter challenges in obtaining real-world graphs for certain topological intervals (e.g., very dense, very large graphs), suggesting that data collection methods must extend beyond simple slicing of dynamic graphs to build a truly comprehensive dataset.

Furthermore, the generalizability of our results is constrained by the demographics of our study participants. While the sample size was appropriate for the experimental design, the group exhibited notable homogeneity in terms of prior experience with graph data. This uniformity potentially limits our insight into how similarity judgments vary across different levels of expertise. For instance, it is unclear whether the overwhelming reliance on *Overall Shape* observed in this study persists among complete novices, or if experts in specific domains might prioritize different topological features.

Throughout our analyses, the independent variables (size, density, layout) often show limited explanatory power regarding the variance in agreement (κ) and correlation (ρ). This may stem from the inherent subjectivity and heterogeneity of human visual perception. However, it is also possible that subtle interaction effects exist but are statistically undetectable due to sample size constraints or participant homogeneity. A larger-scale study with participants from diverse backgrounds would help decouple individual subjectivity from systematic perceptual trends.

From Quantitative to Qualitative Understanding

The primary objective of this study is to quantitatively validate human capabilities in graph similarity assessment and identify computational proxies. While we successfully identify the criteria selected by participants, our current analysis remains primarily descriptive, focusing on reporting observed patterns rather than explaining the cognitive or perceptual foundations underlying these judgments.

For instance, although participants consistently prioritized *Overall Shape*, we haven't theoretically contextualized this finding within established frameworks such as Gestalt psychology or visual attention mechanisms. Consequently, it remains unclear precisely why participants perceive certain global features as salient or how specific geometric properties triggered these decisions. Similarly, for MLLMs, further investigation is needed to link their performance to internal representations, such as attention maps, to fully understand their visual reasoning processes.

To address these gaps, a rigorous qualitative investigation of textual reasoning from both humans and MLLMs is required. By systematically analyzing the generated explanations, future research can identify the specific visual cues that drive similarity judgments and map them to theoretical concepts. Synthesizing this qualitative depth with our quantitative findings will establish a robust theoretical foundation for graph similarity perception in both human and machine agents.

Conclusion

This study presents a comprehensive empirical investigation into the alignment between human visual perception and machine-driven assessments of graph similarity. By constructing a diverse dataset and collecting extensive human judgment data, we establish that human similarity perception is driven primarily by global visual features such

as overall shape and edge density, rather than microscopic topological properties or graph attributes.

Our benchmarking reveals the limitations of traditional computational measures; even the top-performing metric, **Portrait divergence**, captures human perception only moderately. In contrast, our evaluation of state-of-the-art MLLMs highlights a possibility of better alternatives. Models like **GPT-5** not only exhibit superior alignment with human judgments but also offer the distinct advantage of explainability through natural language rationales.

While challenges regarding latency and consistency remain, our findings suggest that MLLMs have the potential to become powerful components in the next generation of visual analytics systems. By serving not only as effective proxies for human perception but also as augmented observers capable of uncovering structural nuances that may elude human eyes, they enable more intuitive, user-centric tools that can both perceive data as analysts do and reveal deeper insights.

Acknowledgements

The authors gratefully acknowledge the support of the BK21 Global Visiting Faculty Fellowship, which facilitated this joint research.

Funding

This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10250156), National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C200520911), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and by the SNU-Global Excellence Research Center establishment project. The ICT at Seoul National University provided research facilities for this study.

Supplemental material

Supplemental material for this article is available online.

Notes

1. Github repository, https://github.com/skwn-j/gsa-dataset
2. Google, gemini-2.5-pro (Latest update: June 2025)
3. OpenAI, gpt-5-2025-08-07
4. Anthropic, claude-sonnet-4-5-20250929

References

1. Assenov Y, Ramírez F, Schelhorn SE et al. Computing topological parameters of biological networks. *Bioinformatics* 2008; 24(2): 282–284. DOI:10.1093/bioinformatics/btm554.
2. Farahani RZ, Miandoabchi E, Szeto WY et al. A review of urban transportation network design problems. *European journal of operational research* 2013; 229(2): 281–302. DOI:10.1016/j.ejor.2013.01.001.
3. Guzman JD, Deckro RF, Robbins MJ et al. An analytical comparison of social network measures. *IEEE Trans Comput Soc Syst* 2014; 1(1): 35–45. DOI:10.1109/TCSS.2014.2307451.
4. Holme P and Saramäki J. Temporal networks. *Physics reports* 2012; 519(3): 97–125. DOI:10.1016/j.physrep.2012.03.001.
5. Moody J, McFarland D and Bender-deMoll S. Dynamic network visualization. *American journal of sociology* 2005; 110(4): 1206–1241. DOI:10.1086/421509.
6. Beck F, Burch M, Diehl S et al. A Taxonomy and Survey of Dynamic Graph Visualization. *Computer Graphics Forum* 2017; 36(1): 133–159. DOI:10.1111/cgf.12791.
7. Kim J et al. DG comics: Semi-automatically authoring graph comics for dynamic graphs. *IEEE Trans Vis Comput Graph* 2024; PP(99): 1–11. DOI:10.1109/TVCG.2024.3456340.
8. Wen X et al. DiffSeer: Difference-based dynamic weighted graph visualization. *IEEE Comput Graph Appl* 2023; 43(3): 12–23. DOI:10.1109/MCG.2023.3248289.
9. Jung S et al. MoNetExplorer: A visual analytics system for analyzing dynamic networks with temporal network motifs. *IEEE Trans Vis Comput Graph* 2023; 30(10): 6725–6739. DOI:10.1109/TVCG.2023.3337396.
10. Jeon H, Quadri GJ, Lee H et al. Clams: A cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE Transactions on Visualization and Computer Graphics* 2024; 30(1): 770–780. DOI:10.1109/TVCG.2023.3327201.
11. Tatu A, Bak P, Bertini E et al. Visual quality metrics and human perception: An initial study on 2D projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*. Roma Italy: ACM. ISBN 978-1-4503-0076-6, pp. 49–56. DOI:10.1145/1842993.1843002.
12. Albuquerque G, Eisemann M and Magnor M. Perception-based visual quality measures. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. pp. 13–20. DOI:10.1109/VAST.2011.6102437.
13. Jardine N, Ondov BD, Elmqvist N et al. The perceptual proxies of visual comparison. *IEEE Trans Vis Comput Graph* 2020; 26(1): 1012–1021. DOI:10.1109/TVCG.2019.2934786.
14. Gleicher M. Considerations for visualizing comparison. *IEEE Trans Vis Comput Graph* 2018; 24(1): 413–423. DOI:10.1109/TVCG.2017.2744199.
15. Quadri GJ and Rosen P. A survey of perception-based visualization studies by task. *IEEE Trans Vis Comput Graph* 2022; 28(12): 5026–5048. DOI:10.1109/TVCG.2021.3098240.
16. Yoghourdjian V et al. Exploring the limits of complexity: A survey of empirical studies on graph visualisation. *Vis Inform* 2018; 2(4): 264–282. DOI:10.1016/j.visinf.2018.12.006.
17. Di Bartolomeo S et al. Evaluating graph layout algorithms: A systematic review of methods and best practices. In *Computer Graphics Forum*. Wiley Online Library, p. e15073. DOI:10.1111/cgf.15073.
18. Herman I, Melancon G and Marshall M. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics* 2000; 6(1): 24–43. DOI:10.1109/2945.841119.
19. Archambault D, Purchase H and Pinaud B. Animation, small multiples, and the effect of mental map preservation in

dynamic graphs. *IEEE Trans Vis Comput Graph* 2010; 17(4): 539–552. DOI:10.1109/TVCG.2010.78.
20. Gao X, Xiao B, Tao D et al. A survey of graph edit distance. *Pattern Anal Appl* 2010; 13(1): 113–129. DOI: 10.1007/s10044-008-0141-y.
21. Weller-Fahy DJ et al. A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Commun Surv Tutor* 2015; 17(1): 70–91. DOI:10.1109/COMST.2014.2336610.
22. Soundarajan S, Eliassi-Rad T and Gallagher B. A guide to selecting a network similarity method. In *2014 SIAM International Conference on Data Mining*. pp. 1037–1045. DOI:10.1137/1.9781611973440.118.
23. Jeon H, Park J, Shin S et al. Stop misusing t-sne and umap for visual analytics, 2025. URL https://arxiv.org/abs/2506.08725. 2506.08725.
24. Jeon H, Lee H, Kuo YH et al. Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25, New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941, pp. 1–24. DOI:10.1145/3706598.3713551.
25. Shneiderman B. *Human-centered AI*. Oxford University Press, 2022.
26. Kim NW, Ahn Y, Myers G et al. How good is chatgpt in giving advice on your visualization design? *ACM Trans Comput-Hum Interact* 2025; 32(5). DOI:10.1145/3745768.
27. Chen N et al. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics* 2024; DOI:10.1109/TVCG.2024.3456320.
28. Choi J, Lee J and Jo J. Bavisitter: Integrating design guidelines into large language models for visualization authoring. In *2024 IEEE Visualization and Visual Analytics (VIS)*. IEEE, pp. 121–125. DOI:10.1109/VIS55277.2024.00032.
29. Rahmanzadehgervi P, Bolton L, Taesiri MR et al. Vision language models are blind. In *Asian Conference on Computer Vision*. pp. 18–34. DOI:10.1007/978-981-96-0917-8_17.
30. Bartolomeo SD, Severi G, Schetinger V et al. Ask and You Shall Receive (a Graph Drawing): Testing ChatGPT's Potential to Apply Graph Layout Algorithms, 2023. DOI: 10.48550/arXiv.2303.08819. 2303.08819.
31. Fan Y, Lyu X, Wang L et al. How well will LLMs perform for graph layout tasks? *Visual Informatics* 2025; : 100285DOI: 10.1016/j.visinf.2025.100285.
32. Miller J, Wallinger M, Felder L et al. Exploring MLLMs Perception of Network Visualization Principles, 2025. DOI: 10.48550/arXiv.2506.14611. 2506.14611.
33. Schetinger V et al. Doom or deliciousness: Challenges and opportunities for visualization in the age of generative models. *Computer Graphics Forum* 2023; 42(2): 423–435. DOI:10.1111/cgf.14841.
34. Huang W, Eades P and Hong SH. Measuring Effectiveness of Graph Visualizations: A Cognitive Load Perspective. *Information Visualization* 2009; 8(3): 139–152. DOI:10.1057/ivs.2009.10.
35. Kypridemou E, Zito M and Bertamini M. The effect of graph layout on the perception of graph properties. *EuroVis (Short Papers)* 2020; DOI:10.2312/evs.20201039.
36. Abbas MM, Aupetit M, Sedlmair M et al. *ClustMe*: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Comput Graph Forum* 2019; 38(3): 225–236. DOI:10.1111/cgf.13684.
37. Gogolou A et al. Comparing similarity perception in time series visualizations. *IEEE Trans Vis Comput Graph* 2018; 25(1): 523–533. DOI:10.1109/TVCG.2018.2865077.
38. Emmert-Streib F, Dehmer M and Shi Y. Fifty years of graph matching, network alignment and network comparison. *Inf Sci (Ny)* 2016; 346-347: 180–197. DOI:10.1016/j.ins.2016.01.074.
39. Donnat C and Holmes S. Tracking network dynamics: A survey using graph distances. *Ann Appl Stat* 2018; 12(2): 971–1012.
40. Masuda N and Holme P. Detecting sequences of system states in temporal networks. *Sci Rep* 2019; 9(1): 795. DOI: 10.1038/s41598-018-37534-2.
41. Wills P and Meyer FG. Metrics for graph comparison: A practitioner's guide. *PLoS One* 2020; 15(2): e0228728. DOI: 10.1371/journal.pone.0228728.
42. Tantardini M, Ieva F, Tajoli L et al. Comparing methods for comparing networks. *Sci Rep* 2019; 9(1): 17557. DOI: 10.1038/s41598-019-53708-y.
43. Koutra D, Vogelstein JT and Faloutsos C. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, pp. 162–170. DOI:10.1137/1.9781611972832.18.
44. Wilson RC and Zhu P. A study of graph spectra for comparing graphs and trees. *Pattern Recognit* 2008; 41(9): 2833–2841. DOI:10.1016/j.patcog.2008.03.011.
45. Sugiyama M, Ghisu ME, Llinares-López F et al. graphkernels: R and python packages for graph comparison. *Bioinformatics* 2018; 34(3): 530–532. DOI:10.1093/bioinformatics/btx602.
46. Yaveroğlu ÖN et al. Revealing the hidden language of complex networks. *Sci Rep* 2014; 4(1): 4547. DOI:10.1038/srep04547.
47. Gleicher M, Albers D, Walker R et al. Visual comparison for information visualization. *Inf Vis* 2011; 10(4): 289–309. DOI:10.1177/1473871611416549.
48. Kerracher N, Kennedy J and Chalmers K. A task taxonomy for temporal graph visualisation. *IEEE Trans Vis Comput Graph* 2015; 21(10): 1160–1172. DOI:10.1109/TVCG.2015.2424889.
49. Zhou H et al. AdaMotif: Graph simplification via adaptive motif design. *IEEE Trans Vis Comput Graph* 2024; DOI: 10.1109/TVCG.2024.3456321.
50. Crnovrsanin T, Chandrasegaran S, Ma KL et al. Staged animation strategies for online dynamic networks. *IEEE Trans Vis Comput Graph* 2020; 27(2): 539–549. DOI:10.1109/TVCG.2020.3030385.
51. Amar R, Eagan J and Stasko J. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, pp. 111–117. DOI:10.1109/INFVIS.2005.1532136.
52. Ghoniem M, Fekete JD and Castagliola P. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE symposium on information visualization*. Ieee, pp. 17–24. DOI:10.1109/INFVIS.2004.1.

53. Guo H, Huang J and Laidlaw DH. Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE transactions on visualization and computer graphics* 2015; 21(10): 1173–1186. DOI:10.1109/TVCG.2015.2424872.
54. Chang C, Bach B, Dwyer T et al. Evaluating perceptually complementary views for network exploration tasks. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. pp. 1397–1407. DOI:10.1145/3025453.3026024.
55. Burch M et al. The state of the art in empirical user evaluation of graph visualizations. *IEEE Access* 2020; 9: 4173–4198. DOI:10.1109/ACCESS.2020.3047616.
56. von Landesberger T et al. Investigating graph similarity perception: A preliminary study and methodological challenges. In *International Conference on Information Visualization Theory and Applications*, volume 4. scitepress.org, pp. 241–250. DOI:10.5220/0006137202410250.
57. Mooney GJ, Purchase HC, Wybrow M et al. The multi-dimensional landscape of graph drawing metrics. In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*. IEEE, pp. 122–131. DOI:10.1109/PacificVis60374.2024.00022.
58. Aupetit M and Sedlmair M. SepMe: 2002 new visual separation measures. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, pp. 1–8. DOI:10.1109/PACIFICVIS.2016.7465244.
59. Lee S, Chang M, Park S et al. Assessing graphical perception of image embedding models using channel effectiveness. In *2024 IEEE Visualization and Visual Analytics (VIS)*. IEEE, pp. 226–230.
60. Quadri GJ and Rosen P. Modeling the influence of visual density on cluster perception in scatterplots using topology. *IEEE Trans Vis Comput Graph* 2021; 27(2): 1829–1839. DOI:10.1109/TVCG.2020.3030365.
61. Xia J et al. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Trans Vis Comput Graph* 2022; 28(1): 529–539. DOI:10.1109/TVCG.2021.3114694.
62. Bae S, Cave K, Görg C et al. Bridging network science and vision science: Mapping perceptual mechanisms to network visualization tasks. *IEEE Trans Vis Comput Graph* 2025; PP. DOI:10.1109/tvcg.2025.3541571.
63. Vázquez PP. Are LLMs ready for visualization? In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*. IEEE, pp. 343–352. DOI:10.1109/PacificVis60374.2024.00049.
64. Basole RC, Major T, Basole RC et al. Generative AI for visualization: Opportunities and challenges. *IEEE Comput Graph Appl* 2024; 44(2): 55–64. DOI:10.1109/MCG.2024.3362168.
65. Park S, Lee S, Choi E et al. Bridging gulfs in ui generation through semantic guidance. *arXiv preprint arXiv:260119171* 2026; DOI:10.48550/arXiv.2601.19171.
66. Park S, Song Y, Lee S et al. Leveraging multimodal llm for inspirational user interface search. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25, New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941, pp. 1–22. DOI:10.1145/3706598.3714213.
67. Bendeck A and Stasko J. An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks. *IEEE Trans Vis Comput Graph* 2024; : 1105–1115DOI:10.1109/TVCG.2024.3456155.
68. Choe K, Lee C, Lee S et al. Enhancing Data Literacy On-Demand: LLMs as Guides for Novices in Chart Interpretation. *IEEE Transactions on Visualization and Computer Graphics* 2025; 31(9): 4712–4727. DOI:10.1109/TVCG.2024.3413195.
69. Sedlmair M and Aupetit M. Data-driven evaluation of visual quality measures. *Comput Graph Forum* 2015; 34(3): 201–210. DOI:10.1111/cgf.12632.
70. Filipov V, Ceneda D, Archambault D et al. Timelighting: Guided exploration of 2d temporal network projections. *IEEE Trans Vis Comput Graph* 2024; DOI:10.1109/TVCG.2024.3514858.
71. Roux J, Bez N, Rochet P et al. Graphlet correlation distance to compare small graphs. *PLoS One* 2023; 18(2): e0281646. DOI:10.1371/journal.pone.0281646.
72. Tsitsulin A, Mottin D, Karras P et al. NetLSD: Hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, pp. 2347–2356. DOI:10.1145/3219819.3219991.
73. Barabási AL and Albert R. Emergence of scaling in random networks. *science* 1999; 286(5439): 509–512. DOI:10.1126/science.286.5439.509.
74. Newman ME and Watts DJ. Renormalization group analysis of the small-world network model. *Physics Letters A* 1999; 263(4-6): 341–346. DOI:10.1016/S0375-9601(99)00757-4.
75. Holland PW, Laskey KB and Leinhardt S. Stochastic blockmodels: First steps. *Social networks* 1983; 5(2): 109–137. DOI:10.1016/0378-8733(83)90021-7.
76. Erdős P and Rényi A. On the existence of a factor of degree one of a connected random graph. *Acta Math Acad Sci Hungar* 1966; 17(359-368): 192. DOI:10.1007/BF01894879.
77. Leskovec J and Sosič R. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2016; 8(1): 1. DOI:10.1145/2898361.
78. Arleo A, Miksch S and Archambault D. Event-based dynamic graph drawing without the agonizing pain. *Comput Graph Forum* 2022; 41(6): 226–244.
79. Hu Y. Efficient, high-quality force-directed graph drawing. *Mathematica journal* 2005; 10(1): 37–71.
80. Doğrusöz U, Madden B and Madden P. Circular layout in the graph layout toolkit. In *International Symposium on Graph Drawing*. Springer, pp. 92–100. DOI:10.1007/3-540-62495-3_40.
81. McInnes L, Healy J and Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426* 2018; DOI:10.48550/arXiv.1802.03426.
82. Sedlmair M, Tatu A, Munzner T et al. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum* 2012; 31(3pt4): 1335–1344. DOI:10.1111/j.1467-8659.2012.03125.x.
83. Pandey AV, Krause J, Felix C et al. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 3659–3669. DOI:10.1145/2858036.2858155.

84. Jain AK. Data clustering: 50 years beyond k-means. *Pattern recognition letters* 2010; 31(8): 651–666. DOI:10.1016/j.patrec.2009.09.011.
85. Huang X, Lai J and Jennings SF. Maximum common subgraph: some upper bound and lower bound results. *BMC bioinformatics* 2006; 7(Suppl 4): S6. DOI:10.1186/1471-2105-7-S4-S6.
86. Ware C. *Information Visualization: Perception for Design*. Interactive Technologies, Morgan Kaufmann, 2019. ISBN 978-0-12-812876-3.
87. Tufte E. The visual display of quantitative information. *Technometrics* 1985; 44: 400–400. DOI:10.1198/tech.2002.s78.
88. Lu M et al. Sticky links: Encoding quantitative data of graph edges. *IEEE Trans Vis Comput Graph* 2024; 30(6): 2968–2980. DOI:10.1109/TVCG.2024.3388562.
89. Nobre C, Zhu K, Mörth E et al. Reading between the pixels: Investigating the barriers to visualization literacy. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. pp. 1–17. DOI:10.1145/3613904.3642760.
90. Xiong C, Van Weelden L and Franconeri S. The curse of knowledge in visual data communication. *IEEE Trans Vis Comput Graph* 2019; 26(10): 3051–3062. DOI:10.1109/TVCG.2019.2917689.
91. Ballweg K, Pohl M, Wallner G et al. Visual similarity perception of directed acyclic graphs: A study on influencing factors and similarity judgment strategies. *Journal of Graph Algorithms and Applications* 2018; 22(3): 519–553. DOI:10.7155/jgaa.00467.
92. Bridgeman S and Tamassia R. A user study in similarity measures for graph drawing. *Graph Algorithms and Applications* 2004; 3: 225–254. DOI:10.1142/9789812796608_0012.
93. Tseng C, Quadri GJ, Wang Z et al. Measuring categorical perception in color-coded scatterplots. In *proceedings of the 2023 CHI conference on human factors in computing systems*. pp. 1–14. DOI:10.1145/3544548.3581416.
94. Berlingerio M, Koutra D, Eliassi-Rad T et al. Netsimile: A scalable approach to size-independent network similarity. *arXiv preprint arXiv:12092684* 2012; DOI:10.48550/arXiv.1209.2684.
95. Bagrow JP and Bollt EM. An information-theoretic, all-scales approach to comparing networks. *Appl Netw Sci* 2019; 4(1). DOI:10.1007/s41109-019-0156-x.
96. Rozemberczki B and Sarkar R. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *29th ACM international conference on information & knowledge management*. pp. 1325–1334. DOI:10.1145/3340531.3411866.
97. Siglidis G, Nikolentzos G, Limnios S et al. GraKeL: A graph kernel library in python. *J Mach Learn Res* 2020; 21(54): 1–5. DOI:10.5555/3455716.3455770.
98. Singh R, Xu J and Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* 2008; 105(35): 12763–12768. DOI:10.1073/pnas.0806627105.
99. Ipsen M. Evolutionary reconstruction of networks. *Func and regulation of cellular systems* 2002; : 241–249DOI:10.1007/978-3-0348-7895-1_23.
100. Ali W, Rito T, Reinert G et al. Alignment-free protein interaction network comparison. *Bioinformatics* 2014; 30(17): i430–i437. DOI:10.1093/bioinformatics/btu447.
101. Heimann M et al. Regal: Representation learning-based graph alignment. In *ACM international conference on information and knowledge management*. pp. 117–126. DOI:10.1145/3269206.3271788.
102. Hermanns J et al. Grasp: Graph alignment through spectral signatures. In *Web and Big Data: 5th International Joint Conference*. Springer, pp. 44–52. DOI:10.1007/978-3-030-85896-4_4.
103. Borgwardt KM and Kriegel HP. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*. IEEE, pp. 8–pp. DOI:10.1109/ICDM.2005.132.
104. Shervashidze N, Schweitzer P, Van Leeuwen EJ et al. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research* 2011; 12(9). DOI:10.5555/1953048.2078187.
105. Landis JR and Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977; : 363–374DOI:10.2307/2529786.
106. Schober P, Boer C and Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia* 2018; 126(5): 1763–1768. DOI:10.1213/ANE.0000000000002864.
107. Tully T, Redfern J, Das D et al. 2025 mid-year llm market update: Foundation model landscape + economics, 2025. URL https://menlovc.com/perspective/2025-mid-year-llm-market-update/. Accessed: 2025-12-22.
108. Zhang Y, Unell A, Wang X et al. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems* 2024; 37: 51727–51753. DOI:10.52202/079017-1639.

Appendix: LLM Prompt

```
# Role
You are an expert researcher in Network Visualization and Graph Drawing. Your task is to
    evaluate visual similarity between node-link diagrams.

# Instruction
I will provide three images of node-link diagrams:
1. **Query Graph (Q)**: The reference graph.
2. **Target Graph 1 (T1)**: The first candidate for comparison.
3. **Target Graph 2 (T2)**: The second candidate for comparison.

**Your Goal:**
Identify which target graph (T1 or T2) is visually more similar to the Query Graph (Q).
    Instead of listing all details, provide a concise justification focusing *only* on the
    decisive factors. You must also **self-evaluate your confidence** in this selection
    based on how distinguishable the similarity is. Finally, quantify the contribution of
    each visual feature.

# Visual Features Definitions
Evaluate the graphs based on the following 6 strictly defined features. The order is fixed
    for the output array.

1. **Overall Structure (Global Topology):** The macro-level shape (e.g., ring, star,
    cluster).
2. **Substructure (Local Patterns):** Recurring small motifs (e.g., triangles, cliques).
3. **Graph Size (Node Count):** Visual estimation of the number of nodes.
4. **Node Degrees (Hubs vs. Leaves):** Distribution of connections (hubs or uniform).
5. **Edge Density (Clutter):** Visual darkness or hairball-likeness.
6. **Number of Communities (Clusters):** Number of visually distinct groups.

# Output Requirements

Step 1. **Internal Evaluation:** Internally compare Q, T1, and T2 across the 6 features to
    determine the winner.
Step 2. **Rationale Formulation:** Formulate a concise explanation that justifies why the
    winner was selected. Directly state the key differences that led to the decision.
Step 3. **Confidence Assessment:** Assign a confidence score (1-5) to your decision based
    on the following scale:
- `1`: **Very confused** (The difference is negligible; almost a random guess).
- `2`: **Confused** (Hard to distinguish, low certainty).
- `3`: **Neutral** (There are differences, but the decision is borderline).
- `4`: **Confident** (The winner is clearly more similar based on visual evidence).
- `5`: **Very confident** (The winner is obviously identical or extremely similar to Q).
Step 4. **Feature Contribution Array:** Provide an array of length 6: [v1: Overall
    Structure, v2: Substructure, v3: Graph Size, v4: Node Degrees, v5: Edge Density, v6:
    Number of Communities].
- `1` (Positive): Winner is *more* similar to Q than the Loser.
- `-1` (Negative): Winner is *less* similar to Q than the Loser.
- `0` (Neutral): Both are equally similar or dissimilar.

# Final Output Format
Please strictly follow the JSON schema provided. The output must be a single JSON object
    with the following fields:
- selected: ['T1' or 'T2']
- rationale: [Concise explanation text]
- confidence: [Integer 1-5]
- features: [Array of 6 integers as defined above]
```