# Scalable Graph Exploration and Visualization: Sensemaking Challenges and Opportunities

Robert Pienta[†], James Abello[*], Minsuk Kahng[†], Duen Horng Chau[†]
[†]Georgia Institute of Technology
[†]{pientars, kahng, polo}@gatech.edu
[*]Rutgers University
[*] abello@dimacs.rutgers.edu

***Abstract*—**

**Making sense of large graph datasets is a fundamental and challenging process that advances science, education and technology. We survey research on graph exploration and visualization approaches aimed at addressing this challenge. Different from existing surveys, our investigation highlights approaches that have strong potential in handling large graphs, algorithmically, visually, or interactively; we also explicitly connect relevant works from multiple research fields – data mining, machine learning, human-computer ineraction, information visualization, information retrieval, and recommender systems – to underline their parallel and complementary contributions to graph sensemaking.**

**We ground our discussion in sensemaking research; we propose a new graph sensemaking hierarchy that categorizes tools and techniques based on how they operate on the graph data (e.g., local vs global). We summarize and compare their strengths and weaknesses, and highlight open challenges. We conclude with future research directions for graph sensemaking.**

## I. Introduction

Making sense of the world is an important part of our everyday lives: researchers want to familiarize themselves with a new field's literature; analysts want to detect suspicious activities before their severity escalate. Gaining insights through making sense of large amounts of data is a fundamental process that advances science, education and technology. In many domains, datasets can often be represented as graphs or networks, as in online social networks (who is connected to whom), network traffic (which computers are communicating), intelligence analysis (who is communicating with whom), and online auctions (who is buying from whom). In this survey, we will use the terms *graph* and *network* interchangeably.

Today, making sense of large graphs remains a fundamental challenge, with few tools that allow users to interactively explore, visualize, and understand million-scale graphs. Data mining and machine learning research has made great strides in developing scalable algorithms, but they typically do not designed to support interactivity or sensemaking tasks. Conversely, principles from human-computer interaction and information visualization that excel in promoting user insight have difficulties scaling to enormously rich information environments with millions of items.

### A. Graph Sensemaking

*Graph sensemaking* refers to the iterative process of understanding and making sense out of graph-formatted data, where a user gradually builds up a representation of the information space to achieve the user's goal [38].

The organizational and cognitive literature has provided a number of theories and models for how people make sense of data (here we focus on graphs). They include Russell et al.'s cost structure view [38], Dervin's sensemaking methodology [39], Klein et al.'s data-frame model [40], and the notional model by Pirolli and Card [41]. While their details may differ, they generally agree that sensemaking is a dynamic, iterative process that involves tasks such as searching, filtering, organizing information, creating schemas, and building evidence.

Furthermore, theories often suggest two predominant sensemaking paradigms: top-down or **global views** and bottom-up or **local views**. Global approaches, best characterized by Shneiderman's mantra "overview, zoom & filter, details-on-demand" pattern in visual information seeking [42], have conventionally received much attention and have worked well for numerous kinds of data in many domains [43], [44], [45], [46], [47], [48], [49], [50], [42]. However, in this big data era, top-down approaches that focus on providing overviews of global information landscapes face significant challenges when applied to graphs with millions or billions of nodes and edges [49], [50]: graph overviews for large graphs are time-consuming to generate [8], [7]; the seminal work on graph clustering by Leskovec & Faloutsos [9] suggests there are simply no perfect overviews (i.e., no single best way to partition graphs into smaller communities), a view echoed by sensemaking literature in that people may have very different mental representations of information depending on their individual goals and prior experiences [51].

Graph sensemaking is a complex and abstract task, highly dependent on both domain and data. For this reason, it is highly unlikely that a single visualization will be sufficient for all sensemaking tasks. In this article, we will cover the works in both global and local paradigms needed for graph sensemaking. We have constructed a *graph sensemaking hierarchy* and placed relevant works in it accordingly (see Figure 1). In our summary and description of numerous approaches, we will focus especially on the **scalability** (both visual and computational) and **interaction techniques** used to improve various aspects of graph sensemaking.

| | | | | |
|---|---|---|---|---|
| Graph Sensemaking | Global View | | Filtering | [1], [2], [3] |
| | | | Sampling | [4], [5], [6] |
| | | | Partitioning | [7], [8], [9], [10] |
| | | | Clustering | [11], [12], [13], [3], [14], [15] |
| | Local View | Free Discovery | Exploration | [16], [17], [14], [18], [3], [15], [19], [20], [21] |
| | | | Network Motifs | [22], [23], [24], [25], [26] |
| | | Targeted Discovery | Pattern Matching | [27], [28], [29], [30], [31] |
| | | | Navigation | [32], [33], [34], [35], [36], [19], [37] |

Fig. 1. **Graph sensemaking hierarchy** of published graph exploration and visualization techniques covered in this survey, organized by their roles. Local views are broken down into *free discovery*, wherein there is no particular known objective, and *targeted discovery*, where the user has a direct data-centric goal.

### *B. Scaling up Graph Sensemaking: Interactive Sensemaking & Scalable Algorithms*

Exploration is a natural first step for a user trying to understand an unfamiliar graph dataset. We distinguish between (1) *exploration*, which is more open-ended, and (2) *navigation*, which has a particular data-centric objective.

With the challenge of exploring an large quantity of data, analysts need a combination of tools to support their abilities and interests [52]. In this survey, we highlight recent research that supports this view, and explore the intricate relationship between scalable algorithms and interaction design of many graph-based techniques. For example, visualization techniques like filtering and data-centric techniques like sampling and approximate pattern finding can work togther to direct expensive computation to much smaller regions of the graph that the user cares about; slow analytics algorithms could hinder interaction, frustrate the user, and reduce discoveries.

**Interactive & Adaptive Views** Consider a user interested in making sense of the genres in a large artist-to-artist music graph. The overall style of the users' investigation may vary in several ways. How different users select new content may be very different [14]; are they interested in a global perspective (e.g. what genres do I listen to most?) or more local perspectives (is there another artist with music like this?) [32].

These questions can be phrased generally for graphs: is the user's investigation mostly detail oriented around certain nodes, or does it concern much larger regions of the graph? Traditional global views offer the user intuition about where their nodes of interest fall in general terms, while local exploration will help the user find detailed information on individual nodes and their neighborhood.

For users browsing locally, they may start looking at different initial genres (e.g. pop vs rock) putting them in different regions of the graph. Among these graph explorers, do most stay in the egonet of where they started or do they traverse larger regions of the graph [33]? Are there coherent patterns in the features of investigated nodes (either for a single user or globally for many)? If so, how can these patterns be leveraged to improve the quality of the tools used in the users' searches [34]? When a single user explores a dataset in search of insight they are traversing a rich data landscape; why not use contemporary machine learning and data mining techniques to better understand *how* and *what* they are looking for during their exploration. Combining information retrieval, machine learning, and data mining with graph exploration is enormously important to answering these questions.

**Scalable Algorithms** Real-time interactivity is crucial to sensemaking. Scalability has been a top priority and success measure in prior and ongoing work [53]. The scalability goals of graph exploration are very different from efforts that aim to visualize the whole graph. Today, it is feasible to lay out a million-node graph, but extreme visual complexity ("hairball") often result [54], suggesting that even if algorithms may run on the whole graph, one may want to only visualize the parts relevant to the user's sensemaking. Recent works have begun to focus on scalable node-local computation, like [55] which can extract a node's 3 million-node egonet in under 150ms.

We are witnessing the birth of interactive systems that integrate scalable machine learning and data mining algorithms with usable user interfaces [14], [56], [57], such as running graph-based inference algorithms (e.g., Belief Propagation) over million-node graphs in sub-second speed in a background thread, keeping the interface responsive. (Recent research reduces that run time even further [58], [59].) More interactive tools can now perform incremental data mining techniques during the user's interaction latency [31]. And some are integrating data-centric approahces such as pre-computation, approximation, and early algorithm termination, by making the systems aware of the I/O and screen bottlenecks and by the careful adoption of such new algorithms [60], [61], [30].

In this survey, we will cover many of the tools, techniques, and contemporary research topics that contribute to graph sensemaking. In Section II, we cover the scalability and techniques needed for graph exploration and visualization. In Section III, we discuss graph interaction techniques and the tools that have pioneered them. Finally we discuss future research directions in Section IV.

## II. Graph Exploration and Visualization

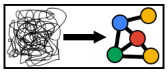
Graph visualization is a challenging area with growing interest spurred by the burgeoning of network datasets. Many tools and techniques have been developed to facilitate discovery; Herman et al. covered much of the initial work in graph visualizations [62]. However, a more recent survey [63] by Landesberger et al. investigated the rich variety of new graph visualization methods since 2000. Both of these works focus primarily on static graphs. The state of the art of *dynamic* graph visualization was recently covered in [64].

Our survey differs from previous surveys both in scope and focus. While our investigation includes works shared among the previous surveys, ours analyzes the scalability and interaction design of the different approaches and tools necessary to the graph sensemaking process (in Section III).

Many graph datasets do not contain spatial node positions, leaving their spatial layout as an exercise for the analyst. Significant research has been done by graph drawing communities investigating how to lay out and summarize entire graphs [65].

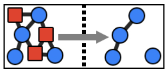
### A. *Visualization & Exploration Techniques*

For graphs that are too large to display at once in full detail, several classes of approaches have been proposed to make them more manageable for exploration. They include sampling, filtering, partitioning, and clustering.



**Graph Sampling** Instead of drawing the entire graph, many approaches sample or filter a subset to reduce a graph's size. Both stochastic and deterministic approaches have been proposed to solve this problem. The stochastic approaches use random sampling techniques to capture a smaller representative graph. A comparison of these approaches and others can be found in [4] and [6].

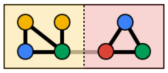
Sampling graphs to maintain their properties is a challenging task especially with scale-free or other power-law graphs. Lee et al. investigated statistical properties of various sampling techniques and found that the method for sampling could heavily bias topological properties like betweeness centrality, assortativity and clustering coefficient [5]. Their work provides important criteria for avoiding biased estimates given different input topologies.



**Graph Filtering** Deterministic sampling and filtering approaches have also been proposed to reduce graph size. Jia et al. have shown that filtering by the approximate betweeness centrality will reduce the graph size while still maintaining the essential structure of the graph [1].

Edges also contribute heavily to visual clutter when drawing a graph. Techniques like *edge bundling* can be used to decrease the amount of space edges take [2].

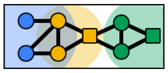
Many networks have such incredible scale that large graph visualizations are not sufficient to perform all exploration tasks. High quality sensemaking for large graphs requires more information than can be summarized at a high level view.



**Graph Partitioning** To improve visual comprehensibility, graph summarization techniques are often used. There are multiple avenues to accomplishing this goal: structural summarization, attribute summarization, and a combination of the two.

A common approach to creating an overview graph is to use partitioning methods on the graph and visualize the partitions [7], [8], [9]. Large graph partitioning is a computationally expensive step and in cases of scale-free and near scale-free networks the partitions may be exceedingly poor [9]. A recent approach called PULP was designed to partition small-world networks and has demonstrated improvements over conventional spectral methods as well as METIS [10].

Many of these methods can scale to the largest of graphs; however, they may take thousands of cores and hours to run. For this reason, partitions are often precomputed and cannot be rerun dynamically during interaction.



**Graph Clustering** Another approach is to create clusters of nodes with similar attributes or to use online analytical processing (OLAP) techniques to roll-up all nodes with a common attribute. In [11], Tian et al. demonstrate *SNAP*; which creates a summary graph by allowing user-specified attributes to determine node-node similarity; and *k-SNAP* which automatically generated subgroups allowing a user to drill-down or roll-up levels of summarization. The *k-SNAP* system works by using OLAP-style aggregation to roll-up multiple nodes by a given attribute, which can be done or undone multiple times, allowing a user to roll-up or drill-down their summary graph. OLAP reduction techniques can be performed quickly for real time systems; however, they do not always produce intuitive reductions.

Combining both structural and attribute information yields a reduced version of the graph where the clusters are both structurally tight and of similar attributes [12], [13]. Zhou et al. proposes a novel distance measure that combines both structural distance as well as node attribute similarity [13]. PivotGraph [12] aggregates nodes and edges based on their attributes; however, it uses a grid-based layout to focus on the relationship between nodes' attributes and connections.

Clusters may also be human-generated, as in [14], [66]. Allowing users to generate and customize their own clusters makes exploration more flexible to changes in input datasets. Rather than relying on a force-directed layout, Schneiderman and Aris propose a static graph layout called *semantic substrates* [66]. Semantic substrates are user-defined, non-overlapping regions in which the nodes are placed according to their attributes. These regions allow users to control edge visibility to provide comprehensibility of each link's source and destination.

### B. *Global and Local Views*

**Global View Challenges** Top-down approaches give an overview of the data by drawing a large, often summarized version of the whole graph dataset. We have analyzed the challenges for top-down graph visualizations:

- Spatial placement of nodes and edges [66]
- Visual incomprehensibility from the number of overlapping nodes and edges [2]
- Representation of node attributes or other node and/or edge information [12], [13]

Often overcoming issues with both visual and computational scalability is a problem for global views. Global views provide high-level information about the structure of a graph.

**Local View Challenges** Local views comprise visualizations in which only a relatively small subset of the entire graph data is shown. In our hierarchy (see Table 1), we taxonomize local views further into those which have an initial objective (Targeted Discovery) and those which are more open-ended (Free Discovery), because their design and interaction are often significantly different.

Because only subgraphs are displayed at a time, these approaches tend to improve visual comprehension, require less computation, but may decrease global insight. A major strength of local views is that algorithms often only need to run on subgraphs, potentially improving scalability. With these strengths, come important challenges:

- at which nodes or subgraphs to start the exploration [67], [57]?
- if a user has a particular goal in mind, can it be predicted by their interactions?
- how will users interact with the system to explore their graph [14]?
- a node may have too many neighbors to draw, can the right nodes be suggested dynamically?

**Free & Targeted Discovery** Graph *navigation* and *exploration* are similar techniques. However, navigation implies that a general destination or objective is known, whereas tools designed for exploration have no such known target. Bottom-up exploration in hierarchical graphs was first investigated in [16] and later expanded on by [17] to incorporate the idea of "degree of interest" to help users identify which nodes to explore. Systems like Apolo [14] do not impose an hierarchy on the data, allowing users to freely define their own clusters, which Apolo incorporates into its machine learning algorithm to infer which nodes the users may want to explore next.

Information retrieval research has focused on scalable approaches to analyze the web-browsing and paths users traversed as they explored the web for millions of users. Click trails have been used to improve the ranking of search results [33], [34]. In many cases, the destinations of such trails can be used directly as search results [35], or even to teleport the user directly to the desired page [36]. Such ideas and techniques may improve the quality of graph exploration tools, yielding more immediately interesting suggestions to the user.

West et al. studied users' abilities and wayfinding techniques as users crawled Wikipedia [32]. They observed a tradeoff wherein users would prefer conceptually simple solutions at the cost of efficiency. In their wayfinding tests, they also investigated how to learn the users' intended targets from their initial movements through Wikipedia pages.

### *C. Subgraph Mining*

Domains from bioinformatics to intelligence analysis often seek particular subgraphs from their data. We taxonomize subgraph mining into two separate areas; pattern matching, in which the user already has some idea of the pattern they seek, and network motif generation, where the common subgraphs are algorithmically detected.

**Pattern Matching** Pattern matching is another bottom-up technique that can aid in exploration tasks, where the user specifies a subgraph of interest (i.e., a graph query) and the matching algorithm looks for simliar instances from a much larger graph (usually called the *data graph*).

At its heart, graph pattern matching is a variation of the subgraph isomorphism problem, an NP-complete task of determining if a given graph is a subgraph of another graph [68]. Exact graph pattern matching is computationally expensive and hard to parallelize. One sensible approach is to search for approximate matchings while another is to leverage special domain attributed subgraphs.

There are a few recent systems that offer approximate subgraph matching, which all focus on large scale techniques. These include MAGE [28], Graphite [69], NeMa [27], TALE [29], and TopKDiv [30]; to name a few. This is essential in scenarios where the user already knows of an interesting pattern exactly or approximately, and wants to find where or how often it occurs in a larger graph.

Many of these systems did not focus on the visualization of the query and results, but rather on the algorithmic and data mining challenges. There is a lot of potential to do background computation during the user's interaction as they build their query. Fan et al. exploited this idea in order to hide the large latencies of graph querying by constructing partial results as a user specified their query pattern [31], [30].

**Frequent Subgraph Mining & Network Motifs** Many works have focused on the discovery of common subgraphs from within much larger graph datasets, which could help discover abnormal activities in the networks, e.g., auction fraud [70], insider trading [71], or insider threats in a company [72]. This explorative process requires almost no foreknowledge of the input graph. Originally coined in [22], [23], *network motifs* are common subgraphs or patterns that occur "unusually" often in a network. Knowing the frequency of subgraphs is not necessarily sufficient to claim that subgraphs are motifs; motifs make a stronger claim by showing that they are statistically more likely in a given input graph than in a random graph of the same size.

Generating the network motifs is a computationally expensive procedure involving subgraph enumeration and aspects of graph similarity from subgraph isomorphism. Although motif detection isn't purely subgraph isomorphism, motif detection approaches currently can detect motifs with dozens of nodes, for modestly sized graphs [25].

The motif mining approach proposed by Milo et al. scans the graph for all $n$-node subgraphs and then compares the occurrence of such $n$-node graphs with their chance to occur in a random-network [23]. Those subgraphs with a statiscally significant appearance rate over the random graphs are considered as motifs. Because this approach scans *all* $n$-node subgraphs it quickly becomes intractable as $n$ increases. Other approaches have improved on the scalability of this work.

Yan et al. created *gSpan*, short for graph-based substructure pattern mining, which discovers frequent graph substructures without the need for a prebuilt candidate list [24]. *gSpan* works by constructing a lexicographic ordering among graphs, which it uses to construct a unique label; it then uses a depth-first search strategy to efficiently mine frequent subgraphs. Grochow et al. further improved motif detection scalability by using subgraph enumeration and symmetry breaking [25]. Symmetry-breaking is a technique by which their algorithm eliminates repeated subgraph isomorphism tests, leading to exponential speedups over the earlier techniques.

Related to motif identification, recent research [26] proposed to develop a vocabulary of common subgraph patterns; (near) cliques, bipartite cores, stars, and others. A graph can then be summarized by replacing the patterns with representative symbols for each pattern, drastically improving visual comprehension.

### *D. Hybrid Graph Visualization*

There are naturally several techniques and tools that combine various graph visualization approaches. By offering several views, a system can overcome the challenges that face a single visualization.

**Overview & Challenges**

- How do we choose an appropriate view given a particular graph [73]?
- How can both structure and lower-level node and edge data be co-visualized? [18], [52]
- How can transitions between views be designed to maximize visualization stability?

Graph visualizations with multiple levels of detail can yield a user improvements in understanding their data, as in [74].

Offering different views also allows easier portrayal of multivariate and other heterogeneous data. This can be seen in [18], where both a graph summary view and a low level multivariate flow chart give users a combination of views. In this approach both the structural behavior of the network as well as the multivariate node data are visualized.

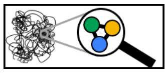
Matrix Zoom [75] is a matrix view strongly coupled with a hierarchical dataview. By offering both a zoomable matrix and a conventional node-link representation, users can smoothly switch to the view that's most useful at their discretion. Henry et al. created NodeTrix [76], a hybrid graph visualization tool designed for social network analysis. NodeTrix uses a connected matrix-view to capture both the sparsity and dense communities often found in social networks.

SocialAction [3] combines structural analysis on graphs and interactive exploration. Their multiple coordinated views contain rankings of nodes for their statistics of structural properties, such as betweenness centrality. These rankings enable users to find interesting nodes systematically and guide them where to start exploration.

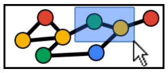
## III. Graph Interaction

Several works have studied the types of common graph interactions. Lee et al. taxonomized common graph visualization interactions in [77]. They separate low level tasks into topological, attribute, and browsing based groups. Here, we will discuss techniques used in graph visualization tools and how they improve graph sensemaking.

**Graph Interaction Basics** User interaction in graph visualizations is essential in all graph exploration tasks. Canonical graph interaction techniques such as brushing, linking, panning, and zooming appear consistently in graph visualizations [78], [79].

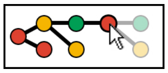


**Lenses** Another common approach is to provide "details on demand" through a simulated lense that provides a detailed view when placed over dense areas. Lenses have proven effective for numerous graph applications [16], [15], [17], [18], [52].



**Graph Selections** There are two main methods for node and edge selection: pointer-selection and query-selection.

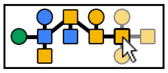
In pointer-selection, the pointer is used to: manually click to select, drag a selection, draw a selection lasso or brush a selection. Query-selection usually uses a query language or filtering interface to let the user specify which nodes they want selected based on node or edge level attributes. Query-selection can be especially useful in cases with rich multivariate node and edge data.



**Structural & Topological Navigation** Topological navigation uses the graphs structure to show and hide portions of the graph based on the connections between nodes. Often this is used so that only a local area of interest is displayed. This is often achieved by only drawing direct neighbors or the *egonet* of a node. This *neighborhood traversal* technique can be very effective means to explore a graph using local topological jumps [19].

TreePlus [21] uses a tree structure to aid in users' exploration of hierarchically clustered graph data. By letting users selectively *grow* the hierarchy, TreePlus strikes a balance between detail and intuition by offering excellent readability, layout stability, and the users' perceptions of tree structure.

In the case of networks with scale-free or near scale-free degree distributions (and other graphs with high degree nodes), pure topological browsing is insufficient, because drawing a single node's neighbors may be drawing a large portion of the graph.



**Degree of Interest Navigation** A more general method than using purely topological information is to use degree of interest to filter. These methods use a degree of interest (DoI) function to hide parts of the graph that are uninteresting to the user. A DoI function evaluates the importance of nodes based on an initial node or group of nodes and produces a ranking for related nodes. Neighborhood traversal can be expressed as a simple DoI function. While the DoI functions proposed originally in [16] used a form of graph distance, other graph-attributes can be used to capture user interest. The initial work on DoI was extended by [17] who show the potential for using other attributes and graph features as inputs to the DoI function. Both of the aforementioned works operated on hierarchies, but could be extended to general graphs.

This eventually gave rise to the notion of tuneable and dynamic DoIs. Abello et al. created a modular DoI for large dynamic networks; wherein they provide the user an interactively defined DoI to improve a user's ability to track critical dynamic elements of their network [37]. The Apolo system integrates machine learning to infer multiple types of DoIs simultaneously [14].

The system *Entourage*, a tool for visualizing biological pathways, uses contextual information provided by the user to visualize interdependencies among pathways [20]. Once a subset of a pathway has been selected, other pathways sharing that subset (or elements from it) are displayed, ranked by their similarity. By leveraging attribute information from the user-selected nodes, Entourage provides interdependencies tuned to the exploration habits of the user.

Modern methods for scalable data mining and machine learning can be used to adapt DoI and other search functionality to a user's tastes.

**Multivariate Graphs** Graph datasets often contain rich metadata in the form of attributes or types on nodes and edges, as graphs can be constructed from any databases with multiple types of entities and relationships [80]. There are a series of work for interacting with this type of multivariate graphs.

PivotPaths [81] shows connecting entities of different types, which allows users to navigate through edges with different semantics. The Entourage system also supports cascades with edges of different types. LinkedVis [82] offers users a visualization of the multivariate graph data. It leverages natural language processing to improve collaborative filtering for job recommendation, but is driven by a user-specified profile containing various entities from LinkedIn's data. Allowing the users to visually generate their own profile and directly effect the recommendation system helps users make sense of their recommendations.

PivotSlice [83] makes users possible to subdivide the entire multivariate data into several parts by constructing a series of dynamic queries.

**Multi-Touch** With the preponderance of mobile and tablet devices, new methods for interfacing with data have arisen. Using multi-touch for rich graph interaction shows great promise.

Schmidt et al. have created a multi-touch interaction set; this set is comprised of TouchPlucking, TouchPinning, TouchStrumming, TouchBundling and PushLens [84]. TouchPlucking allows a user to adjust the placement of edges, TouchPinning fixes a node, TouchStrumming vibrates an edge or all neighboring edges (when a node is strummed), Touchbundling where a user can pinch and bundle edges together, and PushLens which redirects edges around the lense (excluding all nodes and edges within the lens).

While this area is relatively new, the broad variety of input gestures make multi-touch an impressive area for future research. Recent research demonstrating realtime computation on million-scale graphs on an iPad Mini [85], further suggests that such expressive interaction may be combined with scalable data mining algorithms to support large graph exploration on multi-touch devices.

## IV. Research Directions

**Visualization & Exploration Techniques** Graph summarization whether through sampling, filtering, partitioning, clustering, or a combination often requires computationally expensive operations many of which cannot be completed fast enough for a real time system. Rather than relying on precomputation, new methods for graph visualization should investigate faster solutions like iteration or approximation to yield faster summary information (e.g., edge bundling by density estimation [86]).

**Global, Local & Hybrid Views** Global views provide users with an abstract view of their data and yield intuition to their internal models, but alone these may be insufficient for all sensemaking tasks. Many conventional graph visualizations fall into this category. Local views provide a user with low-level, real-time information about their graph data. They trade global insight for greater detail and better scalability. Because of the differences in the way we make sense of graphs, new research will need to investigate how to combine these types of views in ways that balance visual scalability, stability, and comprehensibility. With the right balance, new tools and hybrid visualizations may generalize to a broader variety of research tasks.

**Subgraph Mining** Both graph querying and frequent subgraph mining are incredibly computationally expensive processes, which will likely require approximation to be done in real time. Approximate graph querying lets users ask structured queries of their graph data and receive matches from their data; however, little work has yet to be done on the best ways to portray both the query construction as well as the answer response. A balance must be struck between showing detailed results and providing intuition about where in the graph those results came from.

Network motifs give a user a method to freely explore the common subgraphs in their data. These patterns could be leveraged to greatly improve the visual scalability of the graph, but little research has been done on the best methodology to link the summarized patterns with their reductions in the summary graph.

**Interaction** A plethora of interaction approaches have been designed to improve the ease and efficacy of many graph visualizations. These interaction techniques rely on low latency in order to remain smooth and visually coherent. Many graph algorithms operate too slowly when run on a full graph, for those which cannot be precomputed methods must be developed that can be run iteratively with partial input or quickly approximate based on only a portion of the graph.

With its many natural input gestures *multi-touch* offers new avenues for graph interaction. Multi-touch has numerous challenges to overcome; which gestures are appropriate for which graph actions, which gesture types will work well with graphs, or even which composite actions are meaningful.

**Conclusion** This work has surveyed the state of the art in graph sensemaking and two of its critical components: scalability and interaction design. By building on the work of and drawing research from human-computer-interaction, information visualization, machine learning, data mining, recommendation systems, and information retrieval, the area of graph sensemaking has become a steadily growing research discipline. We produced a graph sensemaking hierarchy consisting of global, local, and hybrid views. We discussed the tools and techniques that contribute to this hierarchy and placed them accordingly. Based on the research we have performed and the literature we have gathered, graph visualization research and the broader area of graph sensemaking research are far from their end. With so many challenging research opporunities, graph sensemaking will continue to inspire new research and vivid graph visualizations in the years to come.

## Acknowledgment

This material is based upon work supported by: NSF Grant No. IIS-1217559, an NSF Graduate Research Fellowship No.

DGE-1148903, and the U.S. ARO and DARPA under Contract Number W911NF-11-C-0088.

## References

[1] Y. Jia, J. Hoberock, M. Garland, and J. Hart, "On the visualization of social and other scale-free networks," *IEEE TVCG*, vol. 14, no. 6, pp. 1285–1292, Nov. 2008.

[2] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE TVCG*, vol. 12, no. 5, pp. 741–748, Sep. 2006.

[3] A. Perer and B. Shneiderman, "Balancing systematic and flexible exploration of social networks," *IEEE TVCG*, vol. 12, no. 5, pp. 693–700, 2006.

[4] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. ACM, 2006, pp. 631–636.

[5] S. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E*, vol. 73, p. 016102, Jan 2006.

[6] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J. H. Cui, and A. G. Percus, "Reducing large internet topologies for faster simulations," ser. NETWORKING'05. Springer-Verlag, 2005, pp. 328–341.

[7] G. Karypis and V. Kumar, "Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0," 1995.

[8] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," in *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3. ACM, 2003, pp. 277–286.

[9] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[10] G. Slota, K. Madduri, and S. Rajamanickam, "Pulp: Scalable multi-objective multi-constraint partitioning for small-world network," in *In the Proceedings of the 2014 IEEE Conference on Big Data*. IEEE, 2014.

[11] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 567–580.

[12] M. Wattenberg, "Visual exploration of multivariate graphs," in *Proceedings of CHI*. ACM, 2006, pp. 811–819.

[13] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 718–729, Aug. 2009.

[14] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos, "Apolo: interactive large graph sensemaking by combining machine learning and visualization," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 739–742.

[15] C. Tominski, J. Abello, F. van Ham, and H. Schumann, "Fisheye tree views and lenses for graph visualization," in *Information Visualization, 2006. IV 2006. Tenth International Conference on*, July 2006, pp. 17–24.

[16] G. W. Furnas, "Generalized fisheye views," *SIGCHI Bull.*, vol. 17, no. 4, pp. 16–23, Apr. 1986.

[17] F. Van Ham and A. Perer, "search, show context, expand on demand: Supporting large graph exploration with degree-of-interest," *IEEE TVCG*, vol. 15, no. 6, pp. 953–960, 2009.

[18] S. van den Elzen and J. J. van Wijk, "Multivariate network exploration and presentation: From detail to overview via selections and aggregations," *IEEE TVCG*, vol. 20, no. 12, pp. 2310–2319, 2014.

[19] J. Heer and D. Boyd, "Vizster: Visualizing online social networks," in *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, ser. INFOVIS '05. IEEE Computer Society, 2005, pp. 5–.

[20] A. Lex, C. Partl, D. Kalkofen, M. Streit, S. Gratzl, A. M. Wassermann, D. Schmalstieg, and H. Pfister, "Entourage: Visualizing relationships between biological pathways using contextual subsets," *IEEE TVCG*, vol. 19, no. 12, pp. 2536–2545, Dec. 2013.

[21] B. Lee, C. S. Parr, C. Plaisant, B. B. Bederson, V. D. Veksler, W. D. Gray, and C. Kotfila, "Treeplus: Interactive exploration of networks with enhanced tree layouts," *IEEE TVCG*, vol. 12, no. 6, pp. 1414–1426, 2006.

[22] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[23] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks." *Science*, vol. 298, no. 5594, pp. 824–827, October 2002.

[24] X. Yan and J. Han, "gspan: graph-based substructure pattern mining," in *ICDM*, 2002, pp. 721–724.

[25] J. A. Grochow and M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology*, ser. RECOMB'07. Springer-Verlag, 2007, pp. 92–106.

[26] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos, "Vog: Summarizing and understanding large graphs," 2014.

[27] A. Khan, Y. Wu, C. C. Aggarwal, and X. Yan, "Nema: Fast graph search with label similarity," *PVLDB*, vol. 6, no. 3, 2013.

[28] R. Pienta, A. Tamersoy, H. Tong, and D. H. Chau, "Mage: Matching approximate patterns in richly-attributed graphs," in *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2014.

[29] Y. Tian and J. Patel, "Tale: A tool for approximate large graph matching," in *ICDE*, 2008.

[30] W. Fan, X. Wang, and Y. Wu, "Diversified top-k graph pattern matching," *PVLDB*, vol. 6, no. 13, 2013.

[31] W. Fan, J. Li, J. Luo, Z. Tan, X. Wang, and Y. Wu, "Incremental graph pattern matching," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '11. ACM, 2011, pp. 925–936.

[32] R. West and J. Leskovec, "Human wayfinding in information networks," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 619–628.

[33] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds: Identifying relevant websites from user activity," in *Proceedings of WWW*, ser. WWW '08. ACM, 2008, pp. 51–60.

[34] A. Singla, R. White, and J. Huang, "Studying trailfinding algorithms for enhanced web search," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10. ACM, 2010, pp. 443–450.

[35] R. W. White and J. Huang, "Assessing the scenic route: Measuring the value of search trails in web logs," in *SIGIR*, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 587–594.

[36] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, "The perfect search engine is not enough: A study of orienteering behavior in directed search," in *Proceedings of CHI*, ser. CHI '04. ACM, 2004, pp. 415–422.

[37] J. Abello, S. Hadlak, H. Schumann, and H. Schulz, "A modular degree-of-interest specification for the visual analysis of large dynamic networks," *IEEE TVCG*, vol. 20, no. 3, pp. 337–350, 2014.

[38] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 1993, pp. 269–276.

[39] B. Dervin, "Information as a user construct: The relevance of perceived information needs to synthesis and interpretation," 1983.

[40] G. Klein, B. M. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives." *IEEE intelligent systems*, vol. 21, no. 4, pp. 70–73, 2006.

[41] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of International Conference on Intelligence Analysis*, vol. 5. Mitre McLean, VA, 2005, pp. 2–4.

[42] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[43] E. H. Chi, "A taxonomy of visualization techniques using the data

state reference model," in *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*. IEEE, 2000, pp. 69–75.

[44] J. Heer and D. Boyd, "Vizster: Visualizing online social networks," in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 2005, pp. 32–39.

[45] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Annual review of information science and technology*, vol. 37, no. 1, pp. 179–255, 2003.

[46] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[47] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: visualizing personal histories," in *Proceedings of CHI*. ACM, 1996, pp. 221–227.

[48] C. Ahlberg and B. Shneiderman, "Visual information seeking: tight coupling of dynamic query filters with starfield displays," in *Proceedings of CHI*. ACM, 1994, pp. 313–317.

[49] D. A. Keim, "Visual exploration of large data sets," *Communications of the ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[50] ——, "Information visualization and visual data mining," *IEEE TVCG*, vol. 8, no. 1, pp. 1–8, 2002.

[51] F. C. Keil, *Concepts, kinds, and cognitive development*. MIT Press, 1992.

[52] C. Tominski, J. Abello, and H. Schumann, "Technical section: Cgv-an interactive graph visualization system," *Comput. Graph.*, vol. 33, no. 6, pp. 660–678, Dec. 2009.

[53] J. Abello, P. M. Pardalos, and M. G. Resende, *Handbook of massive data sets*. Springer, 2002, vol. 4.

[54] G. J. Wills, "Nicheworksinteractive visualization of very large graphs," *Journal of computational and Graphical Statistics*, vol. 8, no. 2, pp. 190–212, 1999.

[55] Z. Lin, M. Kahng, K. M. Sabrin, D. H. P. Chau, H. Lee, and U. Kang, "Mmap: Fast billion-scale graph computation on a pc via memory mapping," in *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2014.

[56] L. Akoglu, D. H. Chau, C. Faloutsos, N. Tatti, H. Tong, J. Vreeken, and L. A. J. V. H. Tong, "Mining connection pathways for marked nodes in large graphs." in *SDM*, 2013, pp. 37–45.

[57] L. Akoglu, D. H. Chau, U. Kang, D. Koutra, and C. Faloutsos, "Opavion: Mining and visualization in large graphs," in *Proceedings of ACM SIGMOD*. ACM, 2012, pp. 717–720.

[58] D. Fisher, I. Popov, S. Drucker *et al.*, "Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster," in *Proceedings of CHI*. ACM, 2012, pp. 1673–1682.

[59] W. G. S. Günnemann, D. Koutra, and C. Faloutsos, "Linearized and single-pass belief propagation," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, 2015.

[60] J. M. Abello and J. S. Vitter, *External Memory Algorithms: DIMACS Workshop External Memory and Visualization, May 20-22, 1998*. American Mathematical Soc., 1999, vol. 50.

[61] J. Abello, F. Van Ham, and N. Krishnan, "Ask-graphview: A large scale graph visualization system," *IEEE TVCG*, vol. 12, no. 5, pp. 669–676, 2006.

[62] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE TVCG*, vol. 6, no. 1, pp. 24–43, 2000.

[63] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, and D. Fellner, "Visual analysis of large graphs: State-of-the-art and future research challenges," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1719–1749, 2011.

[64] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "The state of the art in visualizing dynamic graphs," in *EuroVis - STARs*. Eurographics Association, 2014, pp. 83–103.

[65] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[66] B. Shneiderman and A. Aris, "Network visualization by semantic substrates," *IEEE TVCG*, vol. 12, no. 5, pp. 733–740, 2006.

[67] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2010, pp. 410–421.

[68] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, ser. STOC '71. New York, NY, USA: ACM, 1971, pp. 151–158.

[69] D. H. Chau, C. Faloutsos, H. Tong, J. I. Hong, B. Gallagher, and T. Eliassi-Rad, "Graphite: A visual query system for large graphs," in *ICDM*. IEEE, 2008, pp. 963–966.

[70] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, "Netprobe: a fast and scalable system for fraud detection in online auction networks," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 201–210.

[71] A. Tamersoy, B. Xie, S. L. Lenkey, B. R. Routledge, D. H. Chau, and S. B. Navathe, "Inside insider trading: Patterns & discoveries from a large scale exploratory analysis," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 797–804.

[72] E. Ted, H. G. Goldberg, A. Memory, W. T. Young, B. Rees, R. Pierce, D. Huang, M. Reardon, D. A. Bader, E. Chow *et al.*, "Detecting insider threats in a real corporate database of computer usage activity," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1393–1401.

[73] C. D. Stolper, M. Kahng, Z. Lin, F. Foerster, A. Goel, J. Stasko, and D. H. Chau, "Glo-stix: Graph-level operations for specifying techniques and interactive exploration," *IEEE TVCG*, vol. 20, no. 12, pp. 2320–2328, 2014.

[74] D. Archambault, T. Munzner, and D. Auber, "Topolayout: Multilevel graph layout by topological features," *IEEE TVCG*, vol. 13, no. 2, pp. 305–317, March 2007.

[75] J. Abello and F. van Ham, "Matrix zoom: A visual interface to semi-external graphs," in *Proceedings of InfoVis*. IEEE Computer Society, 2004, pp. 183–190.

[76] N. Henry, J. Fekete, and M. J. McGuffin, "Nodetrix: a hybrid visualization of social networks," *IEEE TVCG*, vol. 13, no. 6, pp. 1302–1309, 2007.

[77] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry, "Task taxonomy for graph visualization," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM, 2006, pp. 1–5.

[78] S. K. Card, J. D. Mackinlay, and B. Shneiderman, Eds., *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., 1999.

[79] C. Ware, *Information Visualization: Perception for Design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.

[80] M. Kahng, S. Lee, and S.-g. Lee, "Ranking objects by following paths in entity-relationship graphs," in *Proceedings of the 4th Workshop for Ph. D. students in information & knowledge management*. ACM, 2011, pp. 11–18.

[81] M. Dork, N. H. Riche, G. Ramos, and S. Dumais, "Pivotpaths: Strolling through faceted information spaces," *IEEE TVCG*, vol. 18, no. 12, pp. 2709–2718, 2012.

[82] S. Bostandjiev, J. O'Donovan, and T. Höllerer, "Linkedvis: exploring social and semantic career recommendations," in *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 107–116.

[83] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, "Interactive exploration of implicit and explicit relations in faceted datasets," *IEEE TVCG*, vol. 19, no. 12, pp. 2080–2089, 2013.

[84] S. Schmidt, M. A. Nacenta, R. Dachselt, and S. Carpendale, "A set of multi-touch graph interaction techniques," in *ACM International Conference on Interactive Tabletops and Surfaces*, ser. ITS '10. ACM, 2010, pp. 113–116.

[85] Y. Chen, Z. Lin, R. Pienta, M. Kahng, and D. H. Chau, "Towards scalable graph computation on mobile devices," in *Proceedings of the 2nd Workshop on Scalable Machine Learning*, 2014.

[86] C. Hurter, O. Ersoy, and A. Telea, "Graph bundling by kernel density estimation," in *Computer Graphics Forum*, vol. 31, no. 3pt1. Wiley Online Library, 2012, pp. 865–874.