Apolo: Interactive Large Graph Sensemaking by Combining Machine Learning and Visualization

Duen Horng “Polo” Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dchau, nkittur, jasonh, christos}@cs.cmu.edu

ABSTRACT

We present APOLO, a system that uses a mixed-initiative approach to help people interactively explore and make sense of large network datasets. It combines visualization, rich user interaction and machine learning to engage the user in bottom-up sensemaking to gradually build up an understanding over time by starting small, rather than starting big and drilling down. APOLO helps users find relevant information by specifying exemplars, and then using a machine learning method called Belief Propagation to infer which other nodes may be of interest.

We demonstrate APOLO’s usage and benefits using a Google Scholar citation graph, consisting of 83,000 articles (nodes) and 150,000 citations relationships. A demo video of APOLO is available at `http://www.cs.cmu.edu/~dchau/apolo/apolo.mp4`.

1. INTRODUCTION

Extracting useful knowledge from large network datasets has become an increasingly important problem in domains ranging from citation networks of scientific literature; social networks of friends and colleagues; and links between web pages in the World Wide Web. Theories of sensemaking provide a way to characterize and address the challenges faced by people trying to organize and understand large amounts of network-based data. Sensemaking refers to the iterative process of building up a representation or schema of an information space that is useful for achieving the user’s goal [13]. For example, a scientist interested in connecting her work to a new domain must build up a mental representation of the existing literature in the new domain to understand and contribute to it. Much existing work on graph visualization aims to develop “summary views” of graphs, on the assumption that the user knows nothing or very little about the data. However, such approaches often struggle to work when faced with large networks with millions of nodes and edges. Sometimes, the approaches simply cannot scale up to such data size due to computation or space requirements; sometimes, the views created are too overwhelming, showing too many nodes and edges; and more often, there are simply no good canonical views that can be generated for the graphs (e.g., real, large social networks often do not have well-defined clusters).

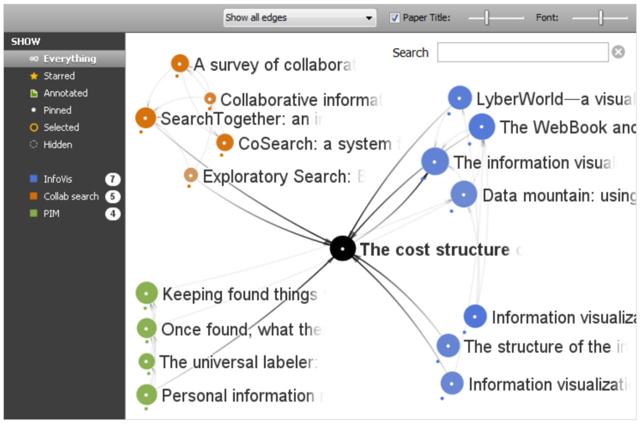


Figure 1: APOLO displaying citation network data around the article *The Cost Structure of Sensemaking*. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs APOLO to find more articles relevant to them.

We have been investigating a different approach where we help the user start with a small subgraph, then incrementally explore and expand, rather than starting big and drilling down. This style of sensemaking can be found in many real-world scenarios. For example, imagine looking for relevant papers for our own research in a citation network; we would often start with papers that we are familiar with, then look at papers that they have cited or being cited by them. We created the APOLO system [4] (Figure 1) to support this kind of bottom-up sensemaking to help the user gradually build up an understanding over time. Within APOLO, a machine learning algorithm called Belief Propagation (BP) guides the user to explore relevant areas of the graph; in turns, the user specifies the nodes that he is interested in as “exemplars”, and BP compute all other nodes’ relevance and show the most relevant ones in the visualization. APOLO provides a number of interaction and visualization features that help people develop and evolve externalized representations of their internal mental models to support sensemaking in large network data. For example, APOLO allows the user to (1) create multiple logical groups (via colors); (2) arrange nodes into spatial groups; and (3) sort nodes by their attributes directly within the visualization for quick comparison.

We summarize our main contributions as follows:

- We aptly select, adapt, and integrate work in machine learning and graph visualization in a novel way to help users make sense of large graphs using a mixed-initiative approach. APOLO goes beyond just graph exploration, and enables users to ex-

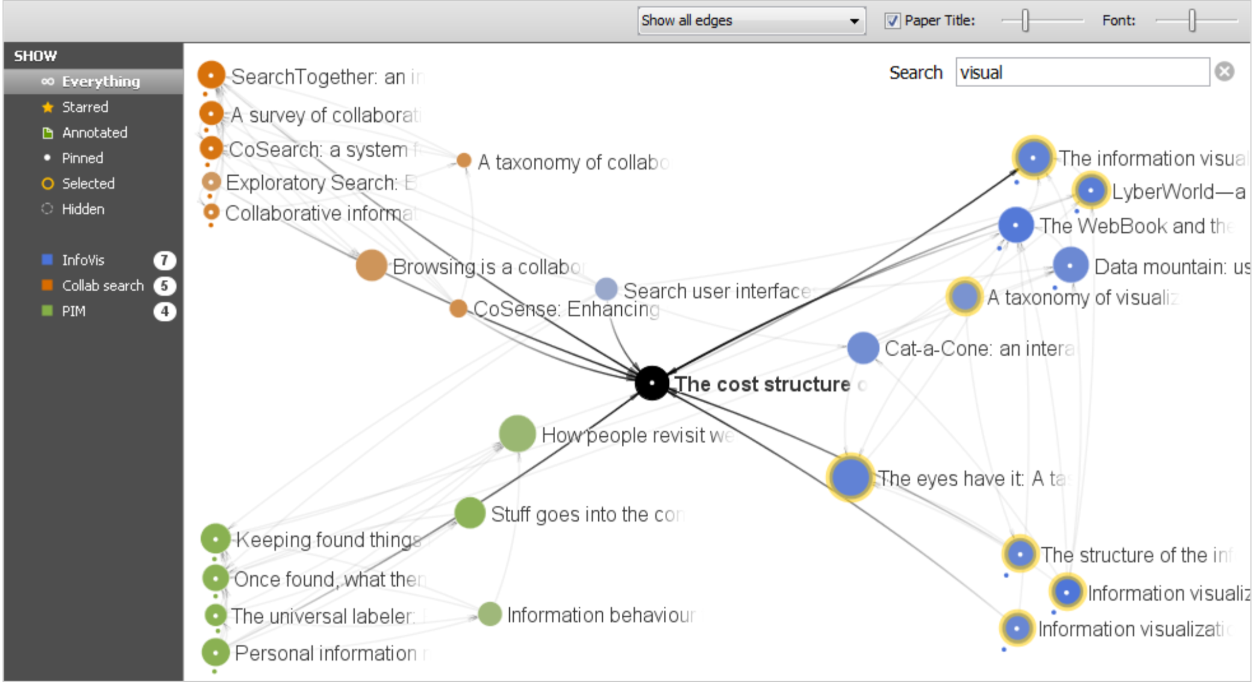


Figure 2: Screenshot of APOLO showing a user exploring the landscape of the related research areas around the article *The Cost Structure of Sensemaking* (highlighted in black). The user gives relevance feedback to APOLO as to which papers he is interested in, as *exemplars* (each with a color dot underneath); and how the exemplars should be categorized (by colors). APOLO incorporates these preferences into BP to find more related nodes, and shows some of the most relevant ones, to avoid overwhelming the user.

ternalize, construct, and evolve their mental models of the graph in a bottom-up manner.

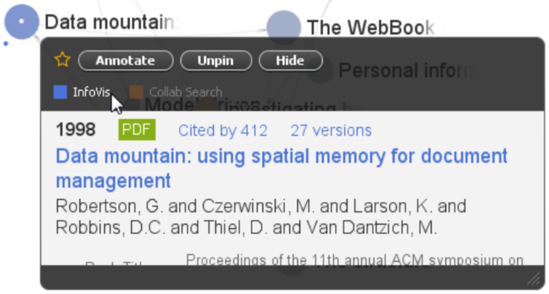
- APOLO offers a novel way of leveraging a complex machine learning algorithm, called Belief Propagation (BP) [15], to intuitively support sensemaking tailored to an individual's goals and experiences; BP was never applied to sensemaking or interactive graph visualization.

2. DEMONSTRATING APOLO

We will demonstrate APOLO's usage, user interaction and fast algorithm through sensemaking scenarios on a Google Scholar citation network dataset, which contains about 83,000 articles (nodes) and 150,000 citations relationships (edges, each represents either a "citing" or "cited-by" relationships).

Scenario. Here, we illustrate one example scenario, where we try to explore and understand the landscape of the related research areas around a seminal article *The Cost Structure of Sensemaking* by Russell et al. (Figure 2 shows final results.) This scenario will touch upon the major features of APOLO. We begin with a single source article highlighted in black in the center of the interface (Figure 4a) and the ten most relevant articles as determined by the built-in BP algorithm. Articles are shown as circles with sizes proportional to their citation count. Citation relationships are represented by directed edges.

After viewing details of an article by mousing over it (Figure 3), the user moves it to a place on the landscape he thinks appropriate, where it remains pinned (as shown by the white dot at the center). The user can also star, annotate, unpin, or hide the node if so desired. After spatially arranging a few articles the user begins to visually infer the presence of two clusters: articles about *information visualization* (*InfoVis*) and *collaborative search* (*Collab Search*). After creating the labels for these two groups (Figure 4a), the user selects a good example article about *InfoVis* and clicks the *InfoVis* label, as shown in Figure 3, which puts the article into that group. A small blue dot (the group's color) appears below the article to indicate it is now an *exemplar* of that group. Changing an article's group membership causes BP to execute and infer the relevance of all other nodes in the network relative to the exemplar(s). A node's relevance is indicated by its color saturation; the more saturated the color, the more likely BP considers the node to belong to the group. Figure 4b-d show how the node color changes as more exemplars are added.



1998 PDF Cited by 412 27 versions

Data mountain: using spatial memory for document management

Robertson, G. and Czerwinski, M. and Larson, K. and Robbins, D.C. and Thiel, D. and Van Dantzich, M.

Proceedings of the 11th annual ACM symposium on

Figure 3: The *Details Panel*, shown when mousing over an article, displays article details obtained from Google Scholar. The user has made the article an *exemplar* the *InfoVis* group by clicking the group's label.

Our user now would like to find more articles for each group to further his understanding of the two research areas. The user right-clicks on the starting paper and selects "Add next 10 most cited neighbors" from the pop-up menu (Figure 5a). By default, new nodes added this way are ranked by citation count (proportional to node size, as shown in Figure 5b) and initially organized in a vertical list to make them easy to identify and process. To see

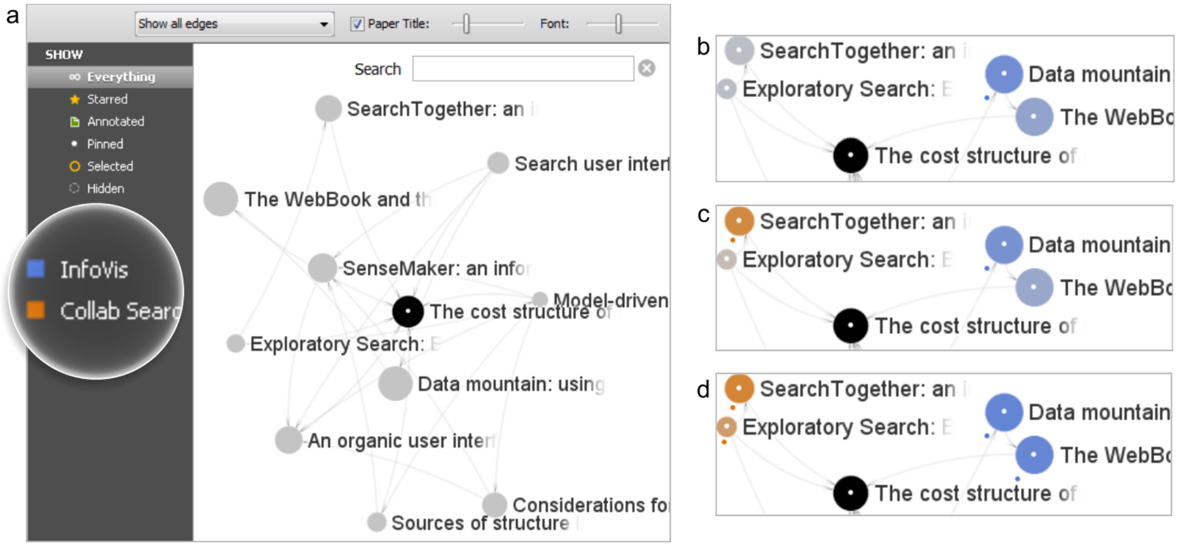


Figure 4: a) In the beginning, the top 10 most relevant articles (in gray) of our starting paper (in black) are shown. Our user has created two groups: *Collab Search* and *InfoVis* (magnified). b-d) Our user first spatially separates the two groups of articles, then assigns them to the groups, making them *exemplars* and causing Apolo to compute the relevance of all other nodes, which is indicated by color saturation; the more saturated the color, the more likely it belongs to the group. The image sequence shows how the node color changes as more exemplars are added.

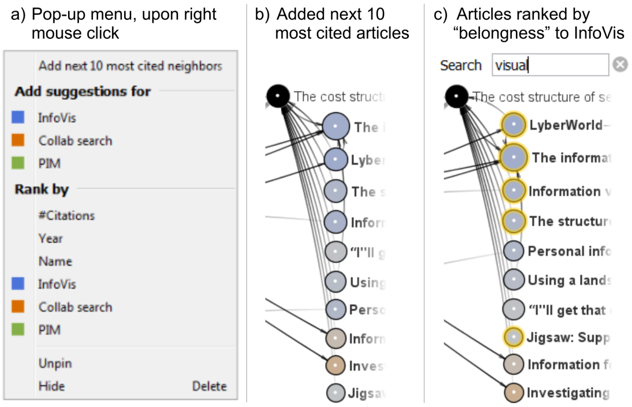


Figure 5: a) Pop-up menu shown upon right click; the user can choose to show more articles, add suggested articles for each group, rank sets of nodes in-place within the visualization, etc. b) Newly added neighboring nodes ranked by citation coun. c) The same nodes, ranked by their "belongness" to the (blue) InfoVis group; articles whose titles contain "visual" are highlighted with yellow halos.

how relevant these new nodes are, he uses Apolo's *rank-in-place* feature to rank articles by their computed relevance to the InfoVis group. To quickly locate the papers about visualization, our user types "visual" in the search box at the top-right corner (Figure 5c) to highlight all articles with "visual" in their titles.

Going further down the list of ranked articles, our users found more InfoVis articles and put them all into that group. Within it, our user further creates two subgroups spatially, as shown in Figure 6, the one on top containing articles about visualization applications (e.g., *Data mountain: using spatial memory for document management*), and the lower subgroup contains articles that seem to provide analytical type of information (e.g., *The structure of the information visualization design space*). Following this work flow, our user can iteratively refine the groups, create new ones, move articles between them, and spatially rearrange nodes in the visualization. The user's landscape of the areas related to Sensemaking following further iterations is shown in Figure 2.

Engaging Our Audience. We will invite our audience to try out APOLO with their own starting set of papers, and collect their feedback on its usability and usefulness. Since we created APOLO to be a general graph sensemaking tool, we will discuss with our audience how APOLO may help with their work in their domains.

3. TECHNICAL DETAILS

The BP algorithm. BP is a fast algorithm; its running time scales linearly with the number of edges in the graph. It has been successfully used in many applications (e.g., error-correcting codes, image de-noising) and domains, but it has never been used for sensemaking. At the high level, BP infers every node's marginal probabilities being in each of the group created by the user. A group's exemplars have high prior probability being in that group, which BP incorporates to perform inference. We implemented BP as described in [10]. The key settings of the algorithm include: (1) a *node potential* function that represents how likely a node belongs to each group (a value closer to 1 means more likely), e.g., if we have two groups, then we assign (0.99, 0.01) to exemplars of group 1, and (0.5, 0.5) to all other nodes; (2) an *edge potential* function that governs to what extent an exemplar would convert its neighbors into the same group as the exemplar (a value of 1 means immediate conversion; we used 0.58).

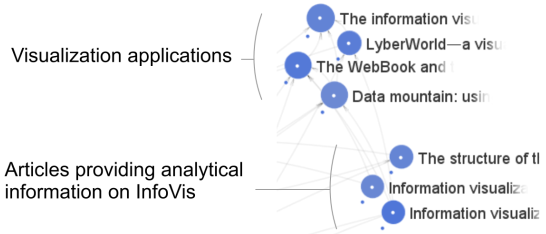


Figure 6: Two spatial subgroups

Implementation. The APOLO system is written in Java 1.6. It uses the JUNG library [11] for visualizing the network. The network data is stored in an SQLite[1] embedded database, for its cross-platform portability. One of our goals is to offer APOLO as a sensemaking tool that work on a wide range of network data, so we designed the network database schema independently from the APOLO system, so that APOLO can readily be used on different network datasets that follow the schema.

4. RELATED WORK

APOLO builds on a large body of research aimed at understanding and supporting how people can gain insights through visualization [8]. APOLO adopts a bottom-up sensemaking approach [13] aimed at helping users construct their own landscapes of information. APOLO provides several features to help people create flexible, ad-hoc groups as they explore, which fit with the process of how they would learn and represent concepts [2]. This "human in the loop" approach differs from a lot of research in graph mining that studies how to automatically discover clusters (or groupings) in graphs, e.g., Graphcut [9], METIS [6].

Much work has been done on developing methods to compute relevance between two nodes in a network; many of them belong to the class of spreading-activation [1, 7, 3]. Used by APOLO, Belief Propagation [15] is a message passing algorithm over link structures similar to spreading activation, but it is uniquely suited for graph sensemaking because it offers *simultaneous* support for: multiple user-specified exemplars (unlike [14]); any number of groups (unlike [1, 5, 14]); linear scalability with the number of edges (best possible for most graph algorithms); and soft clustering, supporting membership in multiple groups (unlike [9]).

Few tools have integrated graph algorithms to interactively help people make sense of network information [5, 12, 14], and they often only support some of the sensemaking features offered by APOLO, e.g., [14] supports one group and a single exemplar.

5. CONCLUSIONS

We present APOLO, a mixed-initiative system for helping users make sense of large network data. Apolo tightly couples large scale machine learning with rich interaction and visualization features to help people explore graphs through constructing personalized information landscapes.

We will demonstrate APOLO's usage and benefits using a Google Scholar citation graph, consisting of 83,000 articles (nodes) and 150,000 citations relationships. We will invite our audience to try out APOLO, comment on its usability and usefulness, and discuss how APOLO may help with their work.

6. ACKNOWLEDGEMENT

This work is partially supported by an IBM Faculty Award, the National Science Foundation under Grants No. CNS-0721736, IIS0970179, IIS-0968484, OCI-0943148 and under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344, and by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the U.S. Government, NSF, or any other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

[1] www.sqlite.org

7. REFERENCES

[1] J. Anderson. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–95, 1983.

[2] L. Barsalou. Ad hoc categories. *Memory & Cognition*, 11(3):211–227, May 1983.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine* 1. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[4] D. Chau, A. Kittur, J. I. Hong, and C. Faloutsos. Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. In *Proceeding of the twenty-ninth annual SIGCHI conference on Human factors in computing systems*. ACM, 2011.

[5] Y. Kammerer, R. Nairn, P. Pirolli, and E. H. Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. CHI*, 2009.

[6] G. Karypis and V. Kumar. METIS: Unstructured graph partitioning and sparse matrix ordering system. *The University of Minnesota*, 2.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[8] B. Kules. From keyword search to exploration: How result visualization aids discovery on the web.

[9] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics*, 22(3):277–286, 2003.

[10] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos. Snare: a link analytic system for graph labeling and risk detection. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1265–1274, New York, NY, USA, 2009. ACM.

[11] J. O'Madadhain, D. Fisher, P. Smyth, S. White, and Y. Boey. Analysis and visualization of network data using JUNG. *Journal of Statistical Software*, 10:1–35, 2005.

[12] A. Perer and B. Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proc. CHI*, pages 265–274, New York, NY, USA, 2008. ACM.

[13] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proc. CHI*, pages 269–276. ACM Press, 1993.

[14] F. van Ham and A. Perer. "Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest. *IEEE TVCG*, 15(6).

[15] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.