

# Overview of Graph Search and Beyond

Omar Alonso  
Microsoft  
Mountain View, CA  
USA

Marti A. Hearst  
UC Berkeley  
Berkeley, CA  
USA

Jaap Kamps  
University of Amsterdam  
Amsterdam  
The Netherlands

## ABSTRACT

Modern Web data is highly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structure, resulting in a massive multidimensional graph. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional IR, and “Graph Search” offers major new ways to access relevant information. Graph search affects both query formulation (complex queries about entities and relations building on the searcher’s context) as well as result exploration and discovery (slicing and dicing the information using the graph structure) in a completely personalized way. This new graph based approach introduces great opportunities, but also great challenges, in terms of data quality and data integration, user interface design, and privacy.

We view the notion of “graph search” as searching information from your personal point of view (you are the query) over a highly structured and curated information space. This goes beyond the traditional two-term queries and ten blue links results that users are familiar with, requiring a highly interactive session covering both query formulation and result exploration. The workshop attracted a range of researchers working on this and related topics, and made concrete progress working together on one of the greatest challenges in the years to come.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Query formulation, Search process, Selection process*

## Keywords

Graph search; Semantic search; Personalization; Exploration; Query suggest

## 1. INTRODUCTION

Information on the Web is increasingly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structure, resulting in a massive multidimensional graph. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional IR, and offers major new ways to access relevant information. In services that rely on personalized information like social networks, the graph plays an even more important role, in other words: *you are the query*.

Graph search affects both query formulation as well as result exploration and discovery. On the one hand, it allows for incrementally expressing complex information needs that triangulate information about multiple entities or entity types, relations between those entities, with various filters on geo-temporal constraints or the sources of information used (or ignored), and taking into account the rich profile and context information of the searcher (and his/her peers, and peers of peers, etc). On the other hand, it allows for more powerful ways to explore the results from various aspects and viewpoints, by slicing and dicing the information using the graph structure, and using the same structure for explaining why results are retrieved or recommended, and by whom.

This new graph based information seeking approach introduces great opportunities, but also great challenges, both technical ranging from data quality and data integration to user interface design, as well as ethical challenges in terms of privacy; transparency, bias and control; and avoiding the so-called filter bubbles. Graph search is already available today in many flavors with different levels of interactivity. Social network-based services like Facebook and LinkedIn provide flexibility to search their personal network from many diverse angles. Web search engines like Google and Bing rely more on using graphs to show related content as a mechanism to include other possible contexts for a given query. Clearly, it is not limited to web, and can be applied to other highly structured data. Just to give an example, the Hansards or parliamentary proceedings are fully public data with a clear graph structure linking every speech to the respective speaker, their role in parliament and their political party. Graph search allows to explore politics from the viewpoint of individual members of parliament or government.

At a high level, graph search seems limited to familiar entity types (e.g., Facebook entities) and templates. How far can this scale? Will this work on truly open domains? There is a huge potential to use the graph to go beyond rec-

ommendations for new friends and contacts or semantically related content. Unlocking the potential of richer knowledge sources for new search strategies requires us to think outside the box, by combining different insights from IR, semantic search, data integration, query expansion and user interfaces to name a few.

The rest of this report is structured in the following way. Section 2 discusses the open research questions raised by searching highly structured data from a personal point of view. Next, in Section 3) we discuss the four keynotes who helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees. Rose Marie Philip talked about personalized post search at Facebook, Swee Lim about graph search at LinkedIn, Doug Oard about good uses for crummy knowledge graphs, and Alex Wade about Microsoft academic graph. Section 4 discusses the six contributed papers, which were presented in a boaster and poster session. Finally, Section 5 provides preliminary discussion of the results and progress made during the workshop.

## 2. OPEN RESEARCH QUESTIONS

We view the notion of “graph search” as searching information from your personal point of view (you are the query) over a highly structured and curated information space. This goes beyond the traditional two-term queries and ten blue links results that users are familiar with, requiring a highly interactive session covering both query formulation and result exploration.

This raises many open questions:

**IR Theory** What happens if search gets personal? Does this break the classic dichotomy between users and documents, as users are nodes in the social network data themselves? What is the consequence of ultimate personalization, as the local graph differs for all users? As the local graph structure is key, does this obviate the need for large central indexes? Do these types of requests fit in the classic paradigm (e.g., Broder’s taxonomy)? How does this shift the balance between the control of the searcher and the ranker over the result set?

**Data Integration** Building a knowledge graph requires massive data integration at many levels: are there trade-offs in simplicity and level of detail (such as the classic knowledge representation trade-off)? What levels of granularity and comprehensiveness are needed for effective deployment? What quality is needed: is any noise acceptable? How to deal with near duplicate detection, conflation, or entity disambiguation?

**Use Cases and Applications** Rather than a universal solution, graph search is particularly useful for specific types of information needs and queries. What are the data and tasks that make graph search works? What kind of scenarios that would benefit from a graph model? In what context can switching perspectives by showing results from the vista of other persons useful?

**Query formulation** How to move from singular queries to highly interactive sessions with multiple variant queries? What new tools are needed to help a searcher construct the appropriate graph search query using refinements or filters to better articulate their needs, or explore fur-

ther aspects? How can we augment query auto-completion to actively prompt user to interactively construct longer queries exploring different aspects?

**Result Exploration** There is a radical shift towards the control of the searcher—small changes in the query can lead to radically different result sets—how can we support active exploration of slices of the data to explore further aspects? Unlike traditional faceted search options, the result space is highly dynamic, how can we provide adaptive exploration options tailored to the context and searcher, at every stage of the process?

**Evaluation** How do we know the system is any good? How to evaluate the overall process, given its personalized and interactive nature? Can we rely on the direct evaluation of query suggestions and query recommendations? Are there suitable behavioral criteria for in the wild testing, such as longer queries, multiple filters, longer dwell-time, more active engagement, more structured-query templates? Can we use standard experimental evaluation methods from HCI and UI/UX design?

**Privacy** Access to personal data is fraught ethical and privacy concerns, is there is similarly structured public data for scientific research? As an extreme form of personalization, how to avoid the uncanny cave, filter bubbles and echo chambers? How ethical is it to privilege a particular query refinement suggestion over the many other possible candidates?

Further discussion on the challenges of graph based search can be found in [1].

## 3. KEYNOTES

Four invited speakers helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees.

### 3.1 Personalized Post Search at Facebook

The opening keynote was given by Rose Marie Philip (Facebook) on “personalized post search at Facebook” [7].

There are over a billion people and over a trillion posts on Facebook. Among these posts, there are uniquely personalized answers to many search queries. The goal of Facebook post search is to help people find the most personally relevant posts for each individual query, tailored to the content of people’s networks. In this talk, I will present some of our work to build a search product that uses personalized graph signals in ranking. I will also give an overview of query modification, posts retrieval and ranking of results.

### 3.2 Graph Search at LinkedIn

The second keynote in the morning was given by Swee Lim (LinkedIn) on “graph search at LinkedIn” [5].

LinkedIn is the largest professional social network. LinkedIn’s graph and search systems help our users discover other users, jobs, companies, schools, and relevant professional information. I will present the evolution of these systems, how they support current use cases, their strengths and weaknesses, our next generation systems, and how we intend to leverage these systems to perform graph searches.

### 3.3 Good Uses for Crummy Knowledge Graphs

The first keynote in the afternoon was given by Doug Oard (University of Maryland) on “good uses for crummy knowledge graphs” [6].

In 1993, Ken Church and Ed Hovy suggested that before we ask how well some new technology meets the need we envision for it, we should pause and first reflect on the question of whether – now that we know something about what can be built – we are envisioning the right uses for what we have. They titled their paper “Good Applications for Crummy Machine Translation” [3]. At about that same time, information retrieval researchers obliged them by (generally without having read their paper) starting to work on cross-language information retrieval; arguably the best application for crummy machine translation ever invented. Now we have some crummy knowledge graphs – and this time we have read the Church and Hovy paper – so perhaps the time is right for us to ask whether we have yet envisioned good uses for crummy knowledge graphs. In this talk, I will seek to seed that discussion.

### 3.4 Overview of Microsoft Academic Graph

The second afternoon keynote was given by Alex Wade (Microsoft Research) with an “overview of Microsoft academic graph” [13]

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals and conference “venues” and fields of study.

## 4. ACCEPTED PAPERS

We requested the submission of short, 4–5 page papers to be presented as booster and poster. We accepted a total of 6 papers out of 8 submissions after peer review (a 75% acceptance rate).

Jadeja and Shah [4] investigate data driven ways to visualize and navigate graph or tree structured data. Navigating or traversing highly curated graph data is an understudied problem, and hierarchical or tree visualizations can help create order and overview. When visualizing the data from the viewpoint of a particular node makes any graph data (such as social network data) look like a tree with the starting node (a person with all context) as point of origin.

Sabetghadam et al. [8] investigates ways of “reranking” results based on a graph traversal approach for multimodal IR, that is hoped to be robust over different distributions of modalities. The use case of multimodal IR in a curated data space with rich context presents a challenge, as different features and scores on different modalities will be very differently distributed in very different probability spaces. An application of Metropolis-Hastings as sampling/estimation method is suggested as (partial) solution.

Sakamoto et al. [9] investigate captioning or summarizing results in highly curated graph data. Succinct descriptions are essential for effective graph exploration, and requires to take the context and structure into account. The paper discusses a particular graph of words, sentences and documents, and also touches upon semantic annotations, which would move the document and text space to an entity space, with all documents and text linked to a particular category or entity.

Santisteban and Cárcamo [10] investigates a variant of the classic Tanimoto or Jaccard similarity measure able to deal with asymmetry in directed graphs and subsumption hier-

archies. Similarity measures are central in IR, and related distance measures central in graph data. The discussion is motivated by a use case of “paradigmatic” structures.

Tong et al. [12] investigate category and word relation graphs for retrieving trouble shooting information/documents, addressing the classical IR problem of human assigned controlled terms versus document free text in the context of a curated data space and rich context (at least in principle). The paper offers an interesting graph approach is outlined, mapping terms to categories, for both requests and documents. Making this graph level explicitly available to users offers interesting new possibilities, and opens up ways to map the noisy term occurrence space to the curated, concept and entity based space of the category codes. Hence this paves the way to a semantic, entity based view.

Yu et al. [14] investigates the strength of connections in an entity graph, specifically a scholarly network with a rich entity graph available as public data. This is an interesting use case with a curated data space and rich context, plus an interesting dynamic structure over time. The paper proposes to take the strength (or weakness) of connections into account — here as simulated blind feedback — turning a network into a weighted network of the simple graph into a valued graph.

## 5. CONCLUSIONS

The workshop brought together researchers from a range of areas in information access, who worked together on searching information from your personal point of view over a highly structured and curated information space. One of the main lines of discussion was the considerable industrial activity around social graphs. The most famous example is Facebook Graph Search, a feature that allows users to perform more sophisticated searches on their social network [11]. Bing has been integrating Facebook into their web search results for the last couple of years. Similarly, Google has been annotating search results with Google+ profiles. And all the rest of the search industry is moving in the same direction.

There are also crucial links with work on searching structured data, and work on the appropriate query languages, in particular as part of semantic search. These branches of research in particular focus on complex querying of structured text or data, whereas the graph search addresses also, and perhaps primarily, the process of constructing series of complex queries interactively. This is directly related to exploratory search and sense making. The graph structure provides natural facets for exploring the data, from a local point of view, allowing for a more dynamic structure than traditional faceted search using rigid, global, hierarchical structure. This challenges our understanding of search user interfaces design and evaluation, with search results moving from the found links, to the HIT page as snippets, and now to query suggestion as previews of possible query extensions.

Graph Search has fundamental consequences for information access and offers tremendous opportunities for building new systems and tools that allow users to explore information from many different angles, shifting control back to the user. This is a radical departure from current systems where the machine learning dominate the interaction: the entire information space is determined by the user, and the user is

in the driver's seat when expressing her needs and exploring the space of options interactive.

**Acknowledgments** This research is funded in part by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014).

## References

- [1] O. Alonso and J. Kamps. Beyond graph search: Exploring and exploiting rich connected data sets. In *ICWE'15: Engineering the Web in the Big Data Era*, volume 9114 of *LNCS*, pages 3–12. Springer, 2015. URL [http://dx.doi.org/10.1007/978-3-319-19890-3\\_1](http://dx.doi.org/10.1007/978-3-319-19890-3_1).
- [2] O. Alonso, M. A. Hearst, and J. Kamps, editors. *GSB'15: Proceedings of the SIGIR'15 Workshop on Graph Search and Beyond*, 2015. CEUR-WS. URL <http://ceur-ws.org/Vol-1393/>.
- [3] K. W. Church and E. H. Hovy. Good applications for crummy machine translation. *Machine Translation*, 8:239–258, 1993. URL <http://dx.doi.org/10.1007/BF00981759>.
- [4] M. Jadeja and K. Shah. Tree-map: A visualization tool for large data. In Alonso et al. [2], pages 9–13. URL <http://ceur-ws.org/Vol-1393/>.
- [5] S. Lim. Graph search at linkedin. In Alonso et al. [2], page 5. URL <http://ceur-ws.org/Vol-1393/>.
- [6] D. W. Oard. Good uses for crummy knowledge graphs. In Alonso et al. [2], page 6. URL <http://ceur-ws.org/Vol-1393/>.
- [7] R. M. Philip. Personalized post search at facebook. In Alonso et al. [2], page 7. URL <http://ceur-ws.org/Vol-1393/>.
- [8] S. Sabetghadam, M. Lupu, and A. Rauber. Leveraging metropolis-hastings algorithm on graph-based model for multimodal ir. In Alonso et al. [2], pages 14–18. URL <http://ceur-ws.org/Vol-1393/>.
- [9] K. Sakamoto, H. Shibuki, T. Mori, and N. Kando. Fusion of heterogeneous information in graph-based ranking for query-biased summarization. In Alonso et al. [2], pages 19–22. URL <http://ceur-ws.org/Vol-1393/>.
- [10] J. Santisteban and J. T. Cárcamo. Unilateral jaccard similarity coefficient. In Alonso et al. [2], pages 23–27. URL <http://ceur-ws.org/Vol-1393/>.
- [11] N. V. Spirin, J. He, M. Develin, K. G. Karahalios, and M. Boucher. People search within an online social network: Large scale analysis of facebook graph search query logs. In *CIKM'14*, pages 1009–1018. ACM, 2014. URL <http://doi.acm.org/10.1145/2661829.2661967>.
- [12] B. Tong, T. Yanase, H. Ozaki, and M. Iwayama. Information retrieval boosted by category for troubleshooting search system. In Alonso et al. [2], pages 28–32. URL <http://ceur-ws.org/Vol-1393/>.
- [13] A. D. Wade. Overview of microsoft academic graph. In Alonso et al. [2], page 8. URL <http://ceur-ws.org/Vol-1393/>.
- [14] Y. Yu, Z. Jiang, and X. Liu. Random walk and feedback on scholarly network. In Alonso et al. [2], pages 33–37. URL <http://ceur-ws.org/Vol-1393/>.