Graph Representations and Applications of Citation Networks

Matthias Petri Alistair Moffat Anthony Wirth
Department of Computing and Information Systems
The University of Melbourne
matthias.petri, ammoffat, awirth@unimelb.edu.au

ABSTRACT

A citation network is a structure of linked documents that share a pool of authors and a pool of subjects, and via citations, provide references to related documents that have preceded them in the chronology of research. In this paper we review citation networks, and survey and categorize the operations that extract data from them. Our goal is to create a framework against which proposed implementations can be assessed, and to provide a basis for research in to algorithms and techniques that might be applied to citation networks. In particular, we seek to extend the concept of "search" over a citation network, to allow for ranked retrieval models in which a wide range of factors influence the list of answers that is presented to the user in response to a query.

1. INTRODUCTION

Eugene Garfield was an early pioneer in the field of bibliometrics and citation analysis, and as early as 1955 had recognized the enormous benefit that could accrue from careful tabulation of the output of scientists – and in particular, that the citations they reported in their published papers were a form of information that could be of benefit to others [9]. Garfield was also an early adopter of computing techniques, and founded the Institute for Scientific Information, or ISI, which in 1964 launched the Science Citation Index, a service that provided systematic coverage of thousands of scientific journals via weekly update listings and annual consolidations, available in hard-copy on a subscription basis to University libraries. Those summaries rapidly became invaluable resources, and allowed researchers to quickly and accurately identify papers that cited their published work, or that were related in other ways. Garfield's vision in this regard also led to the establishment of productivity metrics that allowed the output of researchers and institutions to be tabulated, and created the necessary data for journal impact factors to be defined and computed. Garfield remains active as Emeritus Chairman of Thompson Scientific ISI, and as an author. Indeed, in a recent paper he provides a table of researchers who – identified via ISI data – have been publishing for 69 years or more [11]. He may be planning to occupy a position in a future version of that table.

ADCS'14 , November 27–28 2014, Melbourne, Victoria, Australia

http://dx.doi.org/10.1145/2682862.2682865.

Complementing that early work, in 1967 Ralph Garner proposed the use of graph structures as an adjunct to citation analysis [12]. He introduced his concepts with this:

> The history of science, as all histories, builds on what has come before, and to that extent we can say that science is an edifice built from units or blocks of knowledge which we call scientific papers. When a block of knowledge is added to the structure, we indicate which existing blocks are used to support the new addition, by providing a citation.

The role of citations in acknowledging previous work, and the analysis of them using graph and network structures, remains fundamental to conventional writing and publishing. On-line and electronic publishing structures also retain the same elements, even if employing embedded links that eliminate the explicit bibliography.

The Science Citation Index provided an important resource for searching in the literature associated with some discipline, and ISI and SCI continue as a business. A range of other services are also now available. The most notable of these are, of course, Google Scholar and Microsoft Academic Search, and in the computing disciplines, DBLP and ArnetMiner. Citation networks also apply in less formal domains. The Twitter "re-tweet" can be thought of as a citation link to previous work, as can the Facebook "like".

In this paper we review and survey citation networks, starting with the work of Garfield, and a taxonomy of citation-based search tasks he proposed in 1979. We then consider graph-based citation models, and consider how citation relationships can be efficiently mapped onto manipulable structures amenable to query processing at scale, and with a view to supporting additional querying modalities that step beyond set-based retrieval, and add notions of "locality" and "nearness" to search in citation networks.

2. A CITATION NETWORK MODEL

This section describes the terminology we use to describe a citation network; Table 1 provides an summary of those concepts.

We start with a set $\mathcal{P}$ of n publications interconnected via citations. Each paper $p \in \mathcal{P}$ has associated with it a set of authors, *authors*(p); a set of subjects, *subjects*(p); a set of citations, *predecessors*(p); and a (single) publication forum, *forum*(p). The set of all authors in the citation network is denoted by $\mathcal{A}$; and the set of all subjects appearing in the citation network is denoted by $\mathcal{S}$. Similarly, $\mathcal{F}$ is the set of all forums in which papers in $\mathcal{P}$ have been published. A forum can represent a single conference/year combination, or journal/issue combination, or book. Each paper p in the citation network is published at a specific publication date which is formalized as *pubdate*(p). The set of all papers in $\mathcal{P}$ published by author a is referred to as *papers*(a).

Symbol	Description
$\mathcal{P}$	Set of all publications
$\mathcal{A}$	Set of all authors
$\mathcal{S}$	Set of all subjects
$\mathcal{F}$	Set of all forums
papers(a)	Subset of $\mathcal{P}$ for which a is an author
authors(p)	Subset of $\mathcal{A}$ that are authors of p
coauthors(a)	Subset of $\mathcal{A}$ that are authors of *papers*(a)
predecessors(p)	Subset of $\mathcal{P}$ that are cited by p
predecessors$^*(p)$	Kleene closure of *predecessors*(p)
successors(p)	Subset of $\mathcal{P}$ that cite p
successors$^*(p)$	Kleene closure of *successors*(p)
subjects(p)	Subset of $\mathcal{S}$ that appear in p
keywords(s)	Set of keywords describing $s \in \mathcal{S}$
forum(p)	Forum in $\mathcal{F}$ where p was published
pubdate(p)	Publication date of p
contents(p)	Contents of p, as a sequence of words

Table 1: Notation used to describe a citation network.

Each of the papers $p' \in$ *predecessors*(p) is referenced by p as a part of the context in which the work described in p was carried out. It is usually the case, but not universally so, that if $p' \in$ *predecessors*(p), then *pubdate*$(p') <$ *pubdate*(p): papers are published after the work that they cite. Exceptions occur when authors provide "forward" citations to forthcoming work that has been written and perhaps even accepted, but not yet published. The Kleene closure *predecessors*$^*(p)$ of *predecessors*(p) represents all previous publications that p has built on.

It is also helpful to include the inverse of those relationships, and we refer to all publications $p' \in \mathcal{P}$ that reference a particular paper p as *successors*(p). Similarly, *successors*$^*(p)$ refers to all downstream publications building on the results of p, directly or indirectly. These definitions focus on the citation network as it stands at a given moment in time, with $\mathcal{P}$ fixed and hence *successors*$^*(p)$ finite. However, in a dynamic setting, *successors*$^*(p)$ can be expected to grow as follow-up papers are published. A system that manages citations must thus allow for insert and other update operations, in addition to querying operations.

Each paper p has a set of subjects that categorize its contribution to the scientific literature. These subjects may or may not be explicitly stated in the body of p, and if not explicit, may need to be inferred from the textual content. We regard subjects as being independent concepts that are not required to partition the space of all knowledge; they can have partial overlaps, and "is a" hierarchical relationships if appropriate. So while we might think of "image compression" $\subset$ "data compression", a paper might be about the subject "image compression" as well as about "data compression". To allow definite statements to be made about querying operations, we suppose that each subject s in $\mathcal{S}$ is associated with a unique set of keywords *keywords*(s) that jointly specify that subject, as a categorization fingerprint. With these definitions, there is no one-to-one correspondence between subjects and keywords, as a keyword k might map to multiple distinct subjects: that is, $k \in$ *keywords*(s) and $k \in$ *keywords*(s'), where $s \neq s'$.

Without loss of clarity, we sometimes abuse these definitions, and let *successors*(X), where $X \in \mathcal{P}$ is a set of papers, stand for $\cup_{p \in X}$ *successors*(p). Similarly, *keywords*(p) indicates the union of the set of keywords associated with the subjects in *subjects*(p).

With these definitions, we are able to model sets of scientific papers as a domain that makes use of citations; the next section shows how this formalization addresses information requests.

3. SEARCH IN CITATION NETWORKS

Building a citation network facilitates understanding the flow of ideas, and can help identify experts and useful papers. In this section, we explore typical queries that might be made of citation networks. Indeed, even before computers were prevalent in research institutions, researchers sought (manually curated) citation indexes to digest published research more efficiently, and to stay abreast of developments in specialized areas.

In 1979 Garfield [10, Chapter 5] described a set of ten information needs. The paper-based Science Citation Index (SCI), which he had developed a decade earlier, satisfied most of these needs. Some of Garfield's search categories do not transfer directly to digital citation networks, but most describe search tasks still performed by contemporary researchers. We first revisit a subset of the ten search tasks and examine their applicability within the constraints of the citation network model described above. We then summarize and discuss other search tasks proposed in literature; finally, we consider other search tasks in the context of citation networks.

TASK 1 *Bibliographic Verification Search* [10, p. 42]: Find the paper p from author a published in forum f on subject s.

This is a task that is now almost trivially handled via web search technology, but is nevertheless one that researchers still routinely carry out. A typical scenario would be that we saw a presentation at a conference last year, and know the name of one of the authors and what the paper was about; and seek to obtain full details for this paper so that we can read it carefully, because we can see that it is relevant to our own current activities. Remembering the who and where of a paper is less cognitive overhead than remembering the other publication details.

TASK 2 *Follow-Up Search* [10, p. 47]: Given a publication p, find all $p' \in$ *successors*$^*(p)$ such that *subjects*$(p) \subset$ *subjects*(p').

This task type reflects that it is often of critical interest to continue to trace research activity that builds on the results of an earlier publication p. Publications on the same *subjects*(p) directly citing p will be of interest, as well as subsequent publications $p' \in$ *successors*$^*(p)$ such that *subjects*$(p) \subset$ *subjects*(p').

More than any other, it is this type of search that is associated with citation indexing – the assumption being that publications primarily building on the results of p would explicitly provide a citation to p. Note that this search type does not require that every paper in the chain from p to p' share subjects with p, and it is certainly possible for p' to be a successor of an intervening paper that does not have common subjects with p.

If the subjects of p or a candidate p' are not available, keyword based indicators such as *keywords*$(p') \cap$ *keywords*(p) may be used in the search procedure. Note also that the influence of p on subsequent publications potentially diminishes the more distantly they are connected within the citation network; we return to this notion below.

TASK 3 *Concept Search* [10, p. 50]: Given subject s and a publication p such that $s \in$ *subjects*(p), find all $p' \in \mathcal{P}$ such that $s \in$ *subjects*(p').

Concept Search seeks to identify literature related to a subject s based on a known publication p. Instead of only seeking $p' \in$

successors$^*(p)$ with $s \in$ *subjects*(p') (TASK 2), all publications on the same subject s are of interest. Relevant publications can thus be found by exploring predecessors and successors of p, and widening the search iteratively as more papers are identified. Concept search also requires that papers that satisfy the search criteria but that are not linked to p must also be located. For example, it is not at all uncommon for authors to inadvertently miss citing work that is on the same subject, or work that is removed from the main chronology of a particular discipline area, or published in a language that renders them less accessible. Hence, relying solely on citation linkages from a single published paper is not necessarily a sound strategy. To identify such papers, $A_p =$ *authors*(p) and $f =$ *forum*(p) can also be used to find related publications by exploring *successors*$^*(f)$ or *predecessors*$^*(A_p)$, thereby possibly avoiding the need to traverse the complete network. Note that instead of comparing subjects, *keywords*(s) may be also used to find relevant publications.

TASK 4 *Quick State-of-the-Art Search* [10, p. 57]: Given a time stamp t, and a subject s, find all publications $p \in \mathcal{P}$ such that $s \in$ *subjects*(p) and $t \leq$ *pubdate*(p).

The goal of Quick State-of-the-Art Search is to give the searcher an overview of a specific research area, and avoid the need for them to have to carry out their own more detailed literature review [10]. The publications of interest will include survey publications, or papers that themselves contain multiple citations to papers related to s. Returning recent results (those with *pubdate*$(p) \geq t$) is also of importance as old surveys will not contain coverage of recent developments in the area. Survey publications may be identified by the forum they are published in (for example ACM Computing Surveys) or via their content (for example, the paper title or paper abstract). The number of citations and the length of the publication can also be indicative of survey publications. Publications containing citations to highly cited work or citations to other identifiably survey publications might also be regarded as being suitable matches.

TASK 5 *Disjunctive Keyword Search*: Given a set of keywords K, find all $p \in \mathcal{P}$ such that *keywords*$(p) \cap K \neq \emptyset$.

TASK 6 *Conjunctive Keyword Search*: Given a set of keywords K, find all $p \in \mathcal{P}$ such that *keywords*$(p) \cap K = K$.

Boolean searches are one of the traditional mechanisms for identifying relevant documents. Their drawbacks are also well-known: conjunctive search risks being overly specific, and failing to retrieve desired documents; whereas disjunctive search risks being overly broad, and swamping the user with so many matches that they are unable to effectively extract the subset that is of interest. Library catalog systems of the 1970s and 1980s were typically based on Boolean matching, searching over titles-only, or titles and abstracts. University librarians were respected as the facilitators of search, and were expert at formulating queries constructed as a conjunction of required concepts, with each of those concepts expressed as a disjunction of terms or words.[1]

TASK 7 *Diversity Search*: Given a paper p and a target output size k, return the set of papers $X_k \subset$ *successors*(p) of size at most k that maximizes $|$*subjects*$(X_k) \setminus$ *subjects*$(p)|$.

[1] Title/abstract searches during the 1970s and 1980s were also expensive: one might pay $100 or more for thirty minutes of access time over an acoustic-coupler modem, plus (if in Australia) the cost of the international call to the United States, at several dollars a minute. The result set – a hard-copy listing of abstracts identified by the search – would arrive by mail three days later in the US, or ten days later in Australia. Needless to say, librarians were also respected as the guardians of the institutional search budget.

This task supposes that a researcher may be interested to know of new applications or new interpretations of a familiar work p. The search identifies a set of documents X_k that cite p, and collectively cover a maximally diverse range of subjects. Significant in TASK 7 is that there will thus be a range of topic areas identified where p has some bearing on subsequent work, but where that work does not necessarily share subjects with p. A variant would allow $X_k \subset$ *successors*$^*(p)$ to contribute to X_k, rather than restricting X_k to *successors*(p).

TASK 8 *Expert Search* [3]: Given a set of keywords K, find the authors that have expertise in the set of subjects associated with K.

Expert search seeks to identify experts in a given field, with the searcher providing a set of keywords K that they believe specifies the desired expertise of the authors being sought. An author is regarded as having expertise in a subject s indicated by the keywords K if they have written one or more papers p for which the keywords in K are indicative of subjects $s \in$ *subjects*(p), and if, across the set of papers they have authored, all subjects indicated by K are so covered. Note that the keywords and subjects can also be identified by a set of papers submitted as a query [15].

TASK 9 *Expert Network Search* [6]: Given a set of keywords K, find connected groups of authors such that each group has expertise in all *subjects*(K).

As was the case with TASK 8, Expert Network Search seeks to find authors rather than papers. Now the result can be either a single author (like TASK 8), or a set of collaborating authors that jointly have expertise on all subjects s indicated by K. Intuitively, this search models the process of identifying a research group or other set of closely connected authors that span a set of subjects. Authors might be connected by co-authorship, by common sets of forums they have published in, or by citation of similar papers. Groups might also contains authors that possess none of the desired expertise, but are required to establish connections between experts.

TASK 10 *Recommended Reading* [15]: Given a set of papers X, return a set of additional papers, disjoint from X, that could be of interest.

Küçüktunç et al. [15] suggest that the process of searching for additional papers based on a set of known resources is an important step when new publications are being prepared, to ensure that important citations are not overlooked. The current bibliography X of the draft paper p is used as a query to explore the citation network for further papers of interest. The contents of the papers in X, plus papers in *successors*(X), plus the forums associated with papers in X, plus *contents*(X), might all be factors that are used during the search.

TASK 11 *Recommend Conference or Reviewers* [15]: Given a set of papers X, return a set of forums or authors relevant to X.

The last of the tasks we include in this review is one that immediately precedes the submission of new work for review: that of deciding a forum to send it to, and the optional task of nominating possible referees. Editors and Program Committee Chairs must also find referees for a given paper p that is submitted to their forum. Küçüktunç et al. [15] suggest that these decisions are often made based on the bibliography X of the submitted work – authors evaluate potential forums looking for papers covering subjects similar to those addressed in p, including forums containing papers in X; and editors look for referees who have published in that forum, or who have been cited by p, or who have expertise in the subjects of the paper.

4. HEURISTIC SEARCH

We now extend that set of search tasks, by considering a range of further options.

Sets and sequences. One key change that has taken place in general searching areas is the gradual shift to retrieving an *ordered* list of items, rather than an unordered set. The majority of the search tasks described in Section 2 give rise to a set of answers, and all that is required by a system discharging that task is for the set to be enumerated. Moreover, although the number of papers published is growing rapidly, the bibliography of any given paper will continue to contain, typically, somewhere between ten and a hundred citations. That is, the sizes of sets *predecessors*(p) are unlikely to grow as $\mathcal{P}$ grows, and set-based responses are appropriate for many of the listed tasks.

On the other hand, as the citation network grows in size, some of the task generate answer sets that similarly grow in size. For example, *successors*(p) and *successors*$^*(p)$ might both become large; and a search for related papers via a *Concept Search* (TASK 3) might also generate an unmanageable answer set.

Ranked retrieval methods are now the dominant form of access to large document collections, including the world wide web. In a ranking system, a scoring heuristic is used to estimate the match, or degree of "aboutness", between the query and each document in the collection. The documents are then presented to the users in decreasing score order. Rather than regard the answer as being a set that must be identified in full, a ranking system keeps on offering further documents to the user until they stop asking for them. Hence, such systems are often known as "top-k" mechanisms.

In a citation network, several of the tasks are amenable to a top-k interpretation, with answers presented as the (length-k) prefix of a sequence, rather than as an unordered set. The large answer set generated as a result of evaluating a task query can be converted to a sequence by ordering it using a heuristic score composed a number of contributing factors. This is akin to a web search service ordering documents that are a Boolean match against the supplied query. For example, consider the TASK 2 *Followup Search*, searching for relevant documents in *successors*$^*(p)$. In a citation network, timeliness and recency provide numeric "distance based" components that can be fed in to a weighted average; as do the degree of overlap of subjects and their relative frequencies across the collection; as do the computed academic standing of the authors of the paper being scored, and the academic standing of the forum it was presented in.

Localized searching. Another factor that might usefully be applied to the search tasks is that of localization. As the sizes of both *successors*$^*()$ and *predecessors*$^*()$ increase, it is not feasible to explore the entire citation network at query time. To alleviate this, heuristics that only explore the most promising parts of the network might still lead to interesting results, and execute substantially faster. Dynamic query pruning techniques in information retrieval based on structured inverted indexes allow accelerated query processing without compromising the integrity of the document ranking to some desired depth k; and it should be possible to apply similar techniques to graph search in citation networks, even though the nominal score assigned to an item is a composite of facets, as proposed in the previous paragraph.

Personalization. The search tasks described in Section 3 cover a wide variety of information needs which can be fulfilled by information stored in a citation network. However, often the search agent itself is represented within the citation network as an author or a subject, or even as a paper. Thus, *personalized* search can provide additional benefits to the searcher, provided that appropriate navigational and connectivity cues are made available in the structure storing the network. The paragraphs that follow list a set of features available from the citation network which might usefully add an element of personalization to search tasks, and hence improve the effectiveness of a search system when sequences are being returned rather than sets.

Coauthor Network. As an author in the citation network, all interactions with the citation index occur in the context of the author's own representation in the network. Thus, information "close" to the author or to subjects that the author is familiar with can be seen as being more relevant when search results are being compiled. To a certain extent, these emphases can also be regarded as exerting influence through the various connections in the citation network. Recent work by previous co-authors might be preferenced over papers by more remote connections or by unconnected authors. That is, papers in close proximity to the author's representation in the network are more likely to be relevant to their ongoing research interests.

On the other hand, if the author is keen to find a new point of view, then diversity search is appropriate. That suggests that extending the concept of diversity to also refer to authorship connectivity, and to forums associated with an author, will also be useful. Authors may be very familiar with work published by people that they work with and in the set of forums that they are associated with, and specifically seek to have related work by non-connected authors, or in unexpected forums, drawn to their attention. (Which of us would not want to be informed if a paper we were an author of was cited in *Science* or in *Journal of the ACM*?)

Publication history. The set *papers*(a) of papers published by the author a provides a rich context that can enhance the search effectiveness of the citation index. The expertise of a can be modeled using the union of the papers in *papers*(a). Search results might then include papers chosen according to desired (or the absence of desired) similarity criteria relative to that set.

The last time a has published a paper p on a given subject s provides an indicator of the knowledge of the author at *pubdate*(p). After a long intervening period, without fresh additions to the author's body of work (or at least, not in the area of s), that knowledge can be assumed to have become less relevant to current activities in the area of s. In turn, that could be used to increase the score of papers p' with $s \in$ *subjects*(p') and *pubdate*$(p') >$ *pubdate*(p) when searching for s.

Pseudo-relevance feedback, based on the set of subjects written about by an author over time, *subjects*(*papers*(a)), can also be used to perform query expansion. This option need not be restricted to the author in question, and it might be that relationships between subjects can be mined by considering the citation network as a whole, and looking for commonly co-occurring subject areas across multiple authors.

Past citations. References to other papers in work created by the author, *papers*(a), might help define a broader search scope, and be of direct interest to the searcher. Papers frequently cited by the author can either be given more weight (as they are more important) or less weight (as the author knows about them already). New papers cited by co-authors in other work can be also of value. Citations to *papers*(a) from other authors can also be used to label and categorize *papers*(a). The forums that *papers*(a) is cited in can

indicate new research paths for the author. Linkages might also be gleaned by co-occurrence information. A joint citation at a single location in a paper, "[3,4]", for example, suggests that references "[3]" and "[4]" serve a similar role, a relationship that holds even if neither of "[3]" and "[4]" cite each other.

5. CITATION NETWORK GRAPHS

Garner [12] was the first of many authors to suggest representing citation networks as graphs. Here we discuss the various graph representations that are possible, and consider their applicability to the search tasks defined in the previous sections.

Graph notation. A hypergraph $\mathcal{H} = (V, E)$ consists of a set of (hyper)edges E over a set of vertices V. An edge $(v_i, v_j, \ldots v_p) \in E$ corresponds to an arbitrary subset of vertices in V of size at least two. The arity of the set, p, is the *size* of the edge; in some cases, all hyperedges are required to have the same size. Two vertices $v_i, v_j \in V$ are *adjacent* if there is at least one edge in E containing both v_i and v_j. The *degree* of a vertex corresponds to the number of edges containing it. In some graphs the edges have *weights* associated with them, where the weight of an edge typically represents the cost of using that edge in a path or solution.

Both vertices and edges can additionally have *labels* attached to them, in which case the graph is referred to as being *vertex-labeled* or *edge-labeled* respectively. One example of a labeling is for each vertex or edge to be assigned a color from a small palette; solutions to problems may then be constrained to include only certain colored entities, or to avoid certain color combinations. Typical coloring arrangements constrain adjacent vertices to have distinct colors. If the vertices comprising an edge are ordered and the edge is a sequence rather than a set, the graph is said to be *directed*. Vertices in directed graphs have *in-degrees* and *out-degrees* that need not be equal. A *graph* $\mathcal{G} = (V, E)$ is an instance of a hypergraph in which the size of every edge is two.

A common representation. One standard way of modeling a citation network as a graph $\mathcal{G} = (V, E)$ is to take $V = \mathcal{P}$, the set of papers; and to form a directed edge (p, p') for each $p' \in$ *successors*(p). We regard this graph as modeling edges via a relation R, where $(p, p') \in R$ if $p' \in$ *successors*(p). Batagelj [4] suggests adding two *virtual* vertices v_{-1} and v_n to the graph (assuming that there are n papers, numbered from 0 to $n-1$) to act as common source and sink respectively, to ensure the graph consists of one connected component. Edges are added from v_{-1} (the global source, representing an arbitrarily early paper that any other paper might have cited) to all vertices v_i with in-degree zero; and from each v_j with out-degree zero, to v_n (the global sink, representing a paper written after every other paper, that can potentially cite any of them). The resulting graph is nearly acyclic, as the great majority of the edges are from a paper with an earlier publication time, to a paper with a later publication date.

The edges can also be modeled using the inverse relationship R^{-1} if required, so that they are directed from a node p to the elements in *predecessors*(p). The list $R^{-1}(p, x)$ then represents the papers x that comprise the bibliography of p. As noted earlier, under certain conditions, cycles in citation networks can occur; for example, due to the circulation of journal preprints. Strategies to eliminate these cycles include deleting edges, and duplicating or shrinking cyclic subgraphs [4].

Search operations. Taking a paper-centric view of a citation network allows effective access to citation relationships between papers. Thus, *Follow-Up Search* (TASK 2) can be efficiently resolved via graph traversals by locating the node v_p corresponding to paper p, and, assuming that the edges of the graph represent the relation R, listing the paper associated with the destination vertex of each edge in the out-set of v_p.

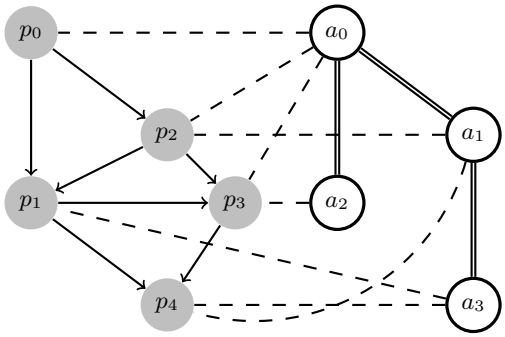

Figure 1: Sample citation network representation with corresponding co-authorship graph. In this example, paper p_1 is cited by p_3 and p_4, and in turn cites p_0 and p_2; and authors a_1 and a_3 are joint writers of paper p_4.

A graph that incorporates both R and R^{-1} can be used to partially perform *Concept Search* (TASK 3) and *Quick State-of-the-Art Search* (TASK 4) by traversing the neighborhood of the vertex v that corresponds to p. However, finding papers related to search concepts which are not closely connected to v_p is more complex in this structure. Carrying out *Recommended Reading Search* (TASK 10) on paper-centric citation graphs has also been shown to be feasible, by performing random walks with restarts (RWR) to compute PageRank metrics for adjacent papers from the set of starting vertices [15]. Küçüktunç et al. [15] also use the set of ranked papers to recommend conferences and reviewers (TASK 11) by accumulating the scores of papers by their conferences and authors.

Given that each paper p has one or more authors, a parallel co-authorship graph can be generated against the underlying citation graph (see Figure 1). Each author is connected to the nodes for all papers they have written or co-written, and pairs of authors are connected by undirected edges if they have co-written any papers. In this structure there are two types (colors) of node, and three types of edge joining them, two of which are undirected.

Garner [12] proposes a different directed author graph representation where edges represent influences between authors. For example, in Figure 1 paper p_2 written by authors a_0 and a_1 is cited by paper p_1 written by author a_3, thus a_0 and a_1 have influenced a_3. Zhou et al. [24] model author relationship based on their social ties, which are deemed to arise when researchers co-author papers, or attend the same conference. This is similar to the two-mode networks described by Wasserman and Faust [23], in which actors interact via events.

Modeling different author relationships enables accessing citation networks from a social or author perspective. If both author and paper relationships are modeled as shown in Figure 1, performing *Bibliographic Verification Search* (TASK 1) by traversing all papers written by an author a_i is efficient. Assuming a set of key papers on a subject has been identified in the graph, the SALSA [16] metric finds putative experts by traversing the edges between author and paper [13]. Identifying *Expert Networks* (TASK 9) can be modeled as finding *Steiner Trees* [6] in the coauthor graph, where expertise is inferred from the subjects associated with the set of papers published by each author [21].

Weights. To model the importance of different relationships, weights can be assigned to edges in the citation network. Thus, weights can help identify experts (TASK 8) or survey papers (TASK 4) by allowing more efficient, greedy explorations of the network. Garner [12] assigns weights to edges based on the number of times paper p_i cites paper p_j. This helps distinguish between what might be termed "core reliance", an innate and critical relationship; and a"passing reference", a brief mention of a standard technique, for example. Hummon and Dereian [14] propose three other edge-weight metrics which measure the importance of an edge by computing the number of times the edge is part of different network exploration algorithms. The "impact" of papers is commonly related to the number of citations which can be modeled by its in-bound links in the graph or by PageRank-like metrics associated with each vertex in the graph [17]. Similarly inferences can be made between the "impact" of a paper p_i, its authors, and the venue c_i it was published in [15, 24]. Citations from papers published in more prestigious venues might be modeled with higher edge weights [20], as might citations by authors who are themselves more highly respected because of their volume of work, or perceived seniority in some other framework. Both of these enhancements allow easier identification of important papers in TASK 4. Walker et al. [22] further incorporate time into the importance of a citation as recent papers are generally cited more often, and unlike the world wide web, citations are fixed and cannot be changed retrospectively.

Aggregate data. Different attributes can be attached to vertices or edges in citation networks. The website ArnetMiner extracts data such as institution and research interests whenever it can identify a personal web page for an author in the network [21]. Similarly each paper vertex in the graph is associated with its contents, *contents*(p). The contents of each paper can then be used to infer expertise of its authors [21]. ObjectRank [2] stores for each unique word in a collection a list of vertices in decreasing "importance" order, allowing efficient access to important papers and experts for each subject. In TASK 3 through to TASK 9, where the searcher supplies query terms describing different concepts, paper and expertise importance ordering is key. This is similar to impact-ordered inverted indexes where postings lists of terms are reordered based on a fixed similarity score [1].

Alternative representations. As suggested by Garner [12], instead of defining the citation relationship $R(p_i, p_j)$ (that is, p_i is cited by p_j) as a matter of simple "existence" or not, it might be helpful to also incorporate the fundamental nature of citations – that they are made in a textual context. A paper p_j that cites paper p_i may contain multiple textual snippets that expose different assessments of p_i. As is now well understood in the context of web search [18], the passage of the text containing the citation (the *anchor* of the citation) can be used as a description of the content of p_i. An example of this representation is shown in Figure 2.

In the example, paper p_8 cites p_5 in two contexts, with text anchors "proposes Y" and "Y uses more space than W". From the labels in the small example, we can infer that p_5 is where technique Y is introduced, that p_6 is where technique X is introduced; and that the authors of p_8 regard X as being an inferior mechanism, so much so, that they didn't even cite the paper it was originally proposed in, and instead have cited a follow-on paper p_7. This edge labeling – provided post-publication by experts in the field, working with the benefit of hindsight – can give insights into the subjects covered by p_i which is valuable for all search tasks described above. The Microsoft academic research search engine[2] provides these anchor texts as "citation contexts" within which a paper is cited. As already noted, citations of papers close to each other for example, "[4,7]" also indicate relationships that might not be explicitly present in the citation network.

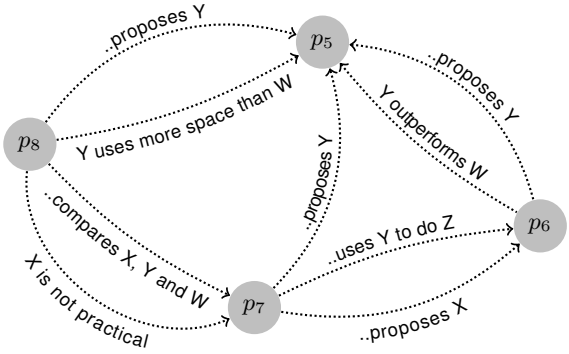

Figure 2: Multiple links between papers augmented with textual anchors, to model the flow of information between papers. Note that these edges are in the R^{-1} space, and are directed from a paper to the works that it cites.

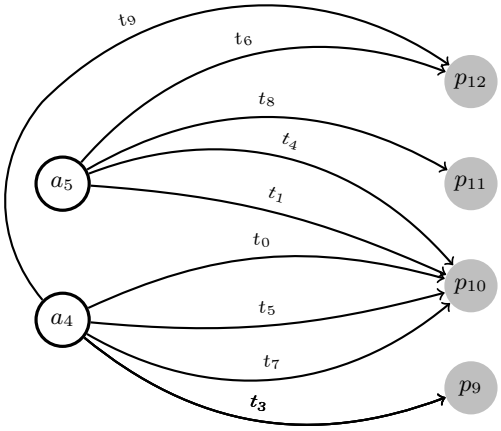

Figure 3: Bipartite time-based author citation graph, modeling authors citing papers at different points in time. In this graph each edge represents an author who, in some unspecified paper written at the time corresponding to the label on the edge, creates a citation to a prior paper.

Instead of modeling relationship between papers, it is also feasible to use graph structures to model the citation history of individual authors. Consider the bipartite graph shown in Figure 3. It shows that author a_4 cites p_{10} in three different papers at time stamps t_0, t_5 and t_7. Time-based graph representations allow viewing a citation network at a given point in time: for example, to answer a query about which papers have been cited by author a_1 prior to time t_7, or when was the first time that author a_5 cited p_{10}. An inverse representation can be constructed showing the time stamps of the citations associated with papers by an author a. These representations allow for interesting variations of different search tasks described in Section 3. Follow-Up Search (TASK 2) is often performed periodically. A searcher might periodically check on recent

[2] http://academic.research.microsoft.com/

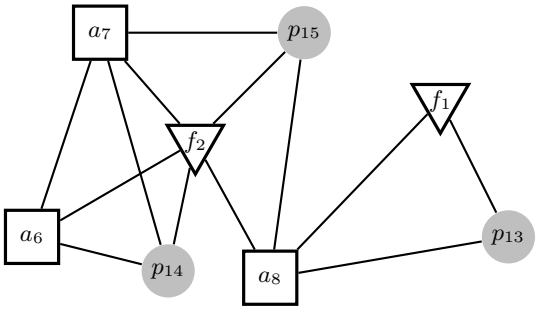

Figure 4: Single citation network graph connecting different entity types.

developments flowing from their work p; instead of all papers citing p, they might only be interested in papers cited since the last time a *Follow-Up Search* was performed.

Adding labels. In general networks, vertices of different types, or "colors" can be employed. Figure 4 gives a small example, where shape is used to denote the label of each vertex. Papers are shaded round nodes; forums are triangular, and authors are square. In the example, authors a_6 and $Author_7$ jointly wrote paper p_{14}, and hence are associated with f_2, that place where that paper was published. This type of representation is similar to the entity-relationship models, commonly used to describe database layouts or semantic web models, where relationships of subject, predicate, object triples are modeled [7, 8]. Information attached to the different node types, such as the contents of a paper, can then propagate through the graph in different ways or using different heuristics. As one example, PopRank [19] uses edge weights to guide how information is passed form node to node. Within this representation, random walks with restarts (RWR) can find experts, conferences or recommended reading. Traversing a graph containing multiple vertex types allows for more complex results and exploration strategies. For example, traversal starting at a paper vertex can lead to forums, authors and other papers to be explored, and returned to the searcher. This is similar to common practices of commercial search engines, which return different types of results (images, news pages, web pages, extracted summaries) to the user for any given query.

Hypergraphs. General hypergraphs [5] can also be used to model citation networks. Taking each vertex to represent an author, each hyperedge represents a single paper that the authors in the hyperedge have co-authored (Figure 5). Similarly, a hypergraph can represent papers and conferences, where each paper is a distinct vertex, and is adjacent via an edge to all other papers in the same conference. Both of these setups has a dual. For example, if each vertex represents a paper, then each can be used to connect the papers authored by a single person.

Directed hypergraphs can be used to model citation relationships between papers, as shown in Figure 6. A hyperedge connects a source vertex representing some paper p with all papers the p cites. Again R^{-1} can be modeled by linking paper p with a single hyperedge to all papers $p' \in \textit{successors}(p)$ which it cites. More complex many-to-many relationships can also be modeled using directed hypergraphs. For example, a paper can be represented as a set of co-authors. Paper p citing paper p' can thus be modeled as a hyperedge between the author sets of p' and p.

Many other such relationships are also possible.

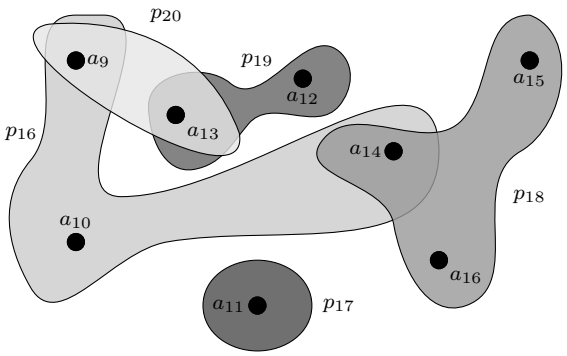

Figure 5: Undirected hypergraph representation of a co-authorship network modeled with authors at the vertices, and papers used to form hyperedges. In this example, paper p_{19} is authored by authors a_{12} and a_{13}.

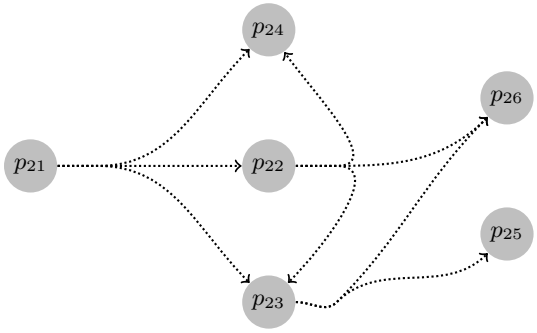

Figure 6: Directed hypergraph representation of a citation network where citation relationships are modeled via directed hyperedges. In this example the hyperedges represent the inverse relation R^{-1}, and the edge out of a vertex leads to the bibliography of that paper.

6. DISCUSSION AND FUTURE WORK

We have reviewed a range of previous work in regard to the representation of citation networks, including search task types, and ways of representing them using various types of graphs, hypergraphs, and weighted or labeled graphs. We have also discussed other extended search tasks that build on the notions of ranking and sequencing, that may be more appropriate for use as citation networks continue to grow, and as the scope implied by any search operation must of necessity become more precise.

The next step is experimentation. We need both (i) data that is realistic of the scientific citation network in terms of both style and scale; and (ii) structures and techniques that can efficiently implement the proposed operations. We are at present exploring our institutional relationship with Springer, with a view to being able to access and mine their extensive data via our library license. We are also exploring the possibility of accessing useful volumes of data from Microsoft Academic Search service. Despite no longer being an active Microsoft project, the interface continues to be available, and provides access to a significant data collection. Finally, while the DBLP collection (`www.dblp.org`) does not include citation information, it provides more than 2.5 million bibliographic records, and may be a useful adjunct to other data, once secured. We will also document the mechanisms and search functionality

offered by current citation networks, including as the ACM Digital Library (`www.acm.org/dl`); Google Scholar (`scholar.google.com`); and ArnetMiner (`www.arnetminer.org`).

Once we have data, we will begin our major project: exploring the tradeoffs that exist between search effectiveness (that is, the quality of the answers that are identified) and search efficiency. At least one full-scale implementation will be constructed and instrumented. We will then be in a position to carry out detailed evaluations on the efficiency and usefulness of advanced search options, noting that the latter might also necessitate the carrying out of a user study.

Acknowledgments. This work was supported under the Australian Research Council's *Discovery Projects* funding scheme (project DP140103256).

References

[1] V. N. Anh and A. Moffat. Pruned query evaluation using pre-computed impacts. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 372–379, 2006. doi: 10.1145/1148170.1148235.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proc. Conf. on Very Large Databases (VLDB)*, pages 564–575, 2004.

[3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 551–558, 2007. doi: 10.1145/1277741.1277836.

[4] V. Batagelj. Efficient algorithms for citation network analysis. *CoRR*, cs.DL/0309023, 2003.

[5] C. Berge. *Graphs and Hypergraphs*, volume 6. North-Holland, 1973. ISBN 9780444103994.

[6] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *Proc. International Conf. on Data Engineering (ICDE)*, pages 431–440, 2002. doi: 10.1109/ICDE.2002.994756.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data: The Story So Far. *International journal on semantic web and information systems*, 5(3):1–22, 2009. doi: 10.4018/jswis.2009081901.

[8] P. P.-S. Chen. The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976. doi: 10.1145/320434.320440.

[9] E. Garfield. Citation indexes for science. *Science*, 122(3159):108–111, 1955. doi: 10.1126/science.122.3159.108.

[10] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, 1979. ISBN 9780894950247.

[11] E. Garfield. The evolution of the Science Citation Index. *International Microbiology*, 10:65–69, 2007. doi: 10.2436/20.1501.01.10.

[12] R. Garner. A computer oriented graph theoretic analysis of citation index structures. In B. Flood, editor, *Three Drexel Information Science Research Studies*. Drexel University Press, 1967.

[13] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. WTF: The who to follow service at Twitter. In *Proc. Conf. on the World Wide Web (WWW)*, pages 505–514, 2013.

[14] N. P. Hummon and P. Dereian. Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11(1):39–63, 1989. doi: 10.1016/0378-8733(89)90017-8.

[15] O. Küçüktunç, E. Saule, K. Kaya, and Ü. V. Çatalyürek. Recommendation on academic networks using direction aware citation analysis. *CoRR*, abs/1205.1143, 2012.

[16] R. Lempel and S. Moran. SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Database Systems*, 19(2): 131–160, Apr. 2001. doi: 10.1016/S1389-1286(00)00034-7.

[17] N. Ma, J. Guan, and Y. Zhao. Bringing PageRank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008. doi: 10.1016/j.ipm.2007.06.006.

[18] O. A. McBryan. GENVL and WWWW: Tools for taming the web. In *Proc. Conf. on the World Wide Web (WWW)*, 1994.

[19] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: Bringing order to web objects. In *Proc. Conf. on the World Wide Web (WWW)*, pages 567–574, 2005. doi: 10.1145/1060745.1060828.

[20] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312, 1976. doi: 10.1016/0306-4573(76)90048-0.

[21] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and mining of academic social networks. In *Proc. Conf. on Knowledge Discovery and Data Mining (WSDM)*, pages 990–998, 2008. doi: 10.1145/1401890.1402008.

[22] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *J. Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007. doi: 10.1088/1742-5468/2007/06/P06010.

[23] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. ISBN 9780521387071.

[24] D. Zhou, S. Orshanskiy, H. Zha, and C. Giles. Co-ranking authors and documents in a heterogeneous network. In *Proc. International Conf. on Data Mining (ICDM)*, pages 739–744, 2007. doi: 10.1109/ICDM.2007.57.