Querying RDF Data from a Graph Database Perspective

Renzo Angles and Claudio Gutierrez

Department of Computer Science, Universidad de Chile
{rangles,cgutierr}@dcc.uchile.cl

Abstract. This paper studies the RDF model from a database perspective. From this point of view it is compared with other database models, particularly with *graph database models*, which are very close in motivations and use cases to RDF. We concentrate on query languages, analyze current RDF trends, and propose the incorporation to RDF query languages of primitives which are not present today, based on the experience and techniques of graph database research.

1 Introduction

The *Resource Description Framework* (RDF) can be viewed from at least two perspectives: (1) From a logical perspective, as a minimal fragment of logic that includes all relevant features needed as representation language for metadata, or as the W3C recommendation [Hay04] says: RDF is an assertional language intended to be used to express propositions using precise formal vocabularies; and (2) From a database perspective, as an extension of data models used in the database community, in particular graph database models. The former point of view has been an active area of research. This does not come as surprise knowing that RDF emerged as a language to represent metadata on the Web, distilling the experience of the community of knowledge representation and Web researchers and developers [LS99]. The latter point of view has received less attention and will be the focus of this paper. We will consider RDF as a data model in the database tradition.

The term *data model* has been used in the database community with different meanings and in diverse contexts. In this paper we will use it in two senses. In a broad or abstract sense, a data model is a collection of conceptual tools for describing the real-world entities to be modeled in the database and the relationships among these entities [SKS96]. In a strict or concrete sense, a *data model*, as defined by Codd [Cod80], is as a combination of three components: (a) a collection of data structure types; (b) a collection of transformation operators and query language and, (c) a collection of general integrity rules.

In the broad sense, RDF can be considered a data model: a collection of conceptual tools for describing real-world entities, namely metadata on the Web. But also in the strict sense of the term, RDF qualifies as well: Point (a) has been more o less addressed at a basic level. One of the documents of the RDF

suite [KC04] speaks of *graph data model* meaning by this concept the data structure implicitly defined by sets of triples. Although one can discuss the precise meaning of this concept [HG04], the graph-like nature of RDF data is clear. Point (c), namely the integrity constraints, is an open issue for RDF, especially when considering the duality of RDF as an open world specification of distributed resources on the Web versus RDF as data model for large single-source repositories with all the issues of a standard database system. The topic of constraints is out the scope of this paper.

This paper concentrates on point (b), namely query languages. This is an active area from a development and implementation point of view, and there is a W3C Working Group addressing the issue of RDF data access, which has a proposal of RDF query language directed mainly to access data on distributed sources [PS04]. There are also works addressing foundational issues, e.g. [HT02,KAC+02,GHO04]. Nevertheless, the discussion about RDF as a full fledged strict database model and the the design and primitives of a query language for such a model is a topic less developed, and probably one of the most needed if we want to take advantage of all the potentialities of the RDF data model (e.g. query optimization, query rewriting, views, update, etc.)

The consideration of RDF as database model puts forward the issue of developing coherently all its database features. In particular its query language should address the kind of queries and problems of the application domains the abstract data model is intended to represent. One of the motivations of this paper are those application domains where interconnection at large scale and navigation of a network is the main modeling theme. Examples of this are biology [Olk03], police-like applications [SAMA+05], navigation in bibliographic databases, etc. To give a flavor of the type of problems, consider queries like: "are suspects A and B related?", submitted to a police database, or "what is the Erdős number of author X", submitted to (a RDF version of) DBLP. The first asks for "relevant" paths connecting these resources in the (RDF) police database, and the second asks simply for the length of the shortest path between the nodes representing Erdős and X. Current proposals of RDF query languages [MKA+02,HBEV04,PS04] do not support[1] queries like these. To address these type of problems, the notions and techniques of *graph databases* can be very valuable. Graph databases are systems designed to support storing and querying information in the form of graphs. They were important, together with object-oriented databases, in the database research of the nineties, and lost part of their appeal after the irruption of semi-structured data models and XML. We claim that graph database models can be a sound support for the design of a RDF database model, particularly for RDF query languages.

Contributions. In this paper we study the RDF model from a database perspective, compare it with other abstract database models, focusing on query languages and graph databases. We restrict in this paper to the logical level,

[1] A language is said to *support* a feature if it provides facilities that make it convenient (reasonable easy, safe and efficient) to use that feature [Str88].

i.e., avoid –when is possible– physical, implementation and indexing considerations. In particular we:

- Compare the RDF model with classical abstract database models putting particular emphasis in graph database models.
- Study current RDF query languages with respect to their capabilities to support graph-like queries and conclude that they give little or no support for them.
- Survey the notions, techniques and systems developed in the area of graph database query languages, and its applicability to the RDF model.
- Propose primitives for RDF query languages based on the graph database experience.

Outline of the paper. Section 2 reviews the motivations and goals of the RDF model and compares it with other database models; Section 3 surveys graph database models and their query languages. Section 4 review current RDF query languages and investigates the support they give for querying graph-like data. Finally, in Section 5 we propose a set of primitives to be incorporated into RDF query languages. Each of them is carefully reviewed against experience of query language development in graph databases.

2 RDF as abstract data model

2.1 Motivations, goals and domains of application.

For the sake of completeness, we briefly recall here the motivations of the designers of RDF, as described in [KC04]: (related models. 1) *Web metadata*, i.e. to provide information about Web resources and the systems that use them (e.g. content rating, capability descriptions, privacy preferences, etc.); (2) Applications that require open rather than constrained information models (e.g. scheduling activities, describing organizational processes, annotation of Web resources, etc.); (3) To do for machine processable information (application data) what the World Wide Web has done for hypertext: to allow data to be processed outside the particular environment in which it was created, in a fashion that can work at Internet scale; (4) Interworking among applications: combining data from several applications to arrive at new information; (5) Automated processing of Web information by software agents: the Web is moving from having just human-readable information to being a world-wide network of cooperating processes. RDF provides a world-wide lingua franca for these processes.

In the same document [KC04] it is stated that the design of RDF is intended to meet the following goals: having a simple data model; having formal semantics and provable inference; using an extensible URI-based vocabulary; using an XML-based syntax; supporting use of XML schema datatypes; allowing anyone to make statements about any resource.

Currently, it is difficult to say which will be the niches of RDF. Although not mentioned in the official documents, the RDF model has impact on other types

of applications as well. Many of these use RDF as a flexible and extensible data model with good support for highly interconnected data. Good examples are the Gene Ontology[2], the Open Directory[3] and the Wordnet Viewer[4].

2.2 Comparison of RDF with other abstract database models

Beginning in the seventies numerous data models have been proposed, each of them with their own concepts and terminology. Surveys and taxonomies of data models are as manifold as data models themselves (see e.g. [SKS96,Nav92,Bee88], [KKT76]). Several of these data models have features relevant for the RDF model. In this section we compare the RDF model with the most important of them. A summary is presented in Table 1.

Physical models. They were the first ones to offer the possibility to organize large collections of data. Among the most important ones are the hierarchical [TL76] and network [TF76] models. These models lack good abstraction level and are very close to physical implementations. The data-structuring is not flexible and not apt to model non-traditional applications. For our discussion they do not much have relevance.

Relational data model was introduced by Codd [Cod83] to highlight the concept of level of abstraction by introducing a clean separation between physical and logical levels. Due to its simplicity of modeling it gained wide popularity among developers and business applications. It is based on the simple mathematical notion of relation, which together with its associated algebra and logic, made the relational model a primary model for database research. In particular, its standard query and transformation language, SQL, became a paradigmatic language for querying.

Although an RDF specification can be logically viewed as a set of binary relations, the differences with the relational model are manifold. Among the most relevant ones are: the relational model was directed to simple record-type data with a structure known in advance (airline reservations, accounting, etc.). The schema is fixed and extensibility is a difficult task. Integration of different schemas is not easy nor automatizable. The query language does not support paths, neighborhoods and queries that address connectivity (an exception is transitivity). There are no objects identifiers, but values.

Semantic models ([PM88]) have their origin in the necessity to provide more expressiveness and incorporate a richer set of semantics into the database from the user point of view. They allow database designers to represent objects and their relations in a natural and clear manner (similar to the way the user view an

[2] `http://www.geneontology.org`
[3] `http://rdf.dmoz.org`
[4] `http://www.openhealth.org/RDDL/wnbrowse`

MODEL	LEVEL	DATA COMPLEX.	CONNECTIVITY	TYPE of DATA
Network	physical	simple	high	homogeneous
Relational	logical	simple	low	homogeneous
Semantic	user	simple/medium	high	homogeneous
Object-O	logical/physical	complex	medium	heterogeneous
XML	logical	medium	medium	heterogeneous
RDF	logical	medium	high	heterogeneous

Table 1: Summary of comparison among different database models. The parameters are: abstraction level, complexity of the data items modeled, degree of connectivity among the data and support to get this information, and finally, flexibility to store different types of data.

application) by using high-level abstraction concepts such as aggregation, classification and instantiation, sub- and super-classing, attribute inheritance and hierarchies [Nav92]. A well-known example is the entity-relationship model [Che76]. It has become a basis for the early stages of database design, but due to lack of preciseness cannot replace models like relational or O-O.

For RDF database research, semantic models are relevant because they are based on a graph-like structure which highlights the relations between the entities to be modeled.

Object oriented data models ([Kim90]) are based on the object-oriented programming paradigm. Their objective is representing data as a collection of objects that are organized in classes and have complex values and methods associated with them. They are intended to model non-conventional database applications consisting of complex objects systems with many semantically interrelated components as in CAD/CAM, computer graphics or information retrieval.

Object-oriented database models have been related to Graph database ones because the explicit or implicit graph structure in their definitions [LP91], [AGP$^+$92], [GPdBG90]. Nevertheless, there remain important differences rooted in the form that each of them models the world. O-O models view the world as a set of objects having certain state (data) and interacting among them by methods. On the contrary, graph database models, and RDF in particular, model the world as a network of relations. The emphasis in RDF is on the interconnection of the data, the network of relations among the data and the properties of these relations. The emphasis of O-O is on the objects, their values and methods. However, there are proposals to apply O-O concepts to RDF [BV04].

Semistructured data models ([Bun97,AQM$^+$97,Abi97]) are oriented to model semi-structured data. Of all the most visible models in the literature, the semi-structured data model is one of the closest in several points to RDF. Semi-structured models deal with data whose structure is irregular, implicit and partial, and whose schema is usually very large, contained within the data itself, and rapidly evolving [Abi97]. One of the best representative is OEM [PGMW95].

It is a model based on objects, which have unique identifiers, and values that can be simple types or object references. There is a natural graph representation: objects are nodes, and values are labeled arcs. The main differences with RDF are: the lightweight inferencing available, the existence of blank nodes, the stronger typing system and the fact that labels are also nodes in RDF.

Another representative is the XML model [BPSM]. There are substantial differences between XML and RDF. First, RDF has a higher abstraction level; in fact RDF is an application of XML to represent metadata. Structurally XML has a ordered-tree-like structure against the graph structure of RDF. At the semantic level, in XML the information about the data is part of the data (in other words XML is self-describing); in contrast, RDF expresses explicitly the information about the data using relations between entities. An important advantage of RDF is its extensibility in both schema and instance level. See [GR02,ADL$^+$04] for a major comparison of these models.

3 Graph database Models and their query languages

3.1 Graph database models

Graph database models appeared with the objective of modeling information whose logical structure is a graph. In this sense, they are the closest to the RDF model by the data type used. Among the first ones, we have the the Logical Data Model [KV84,KV93] and the Functional Data Model [Shi81], which define an implicit structure of labeled graphs. The Logical data model introduces basic, composition, and collection nodes, all of which can be modeled in RDF. On the other hand, in many semantic and object oriented data models the conceptual representation of data is transparently graph-based. For example O_2 [LRV88] defines basic, tuple-structured, and set-structured types (the first type is similar to RDF blank node and the remainder two can be modeled as relations in RDF); GOOD [GPdBG90] is oriented primarily to graphical user interfaces; OEM [PGMW95] addresses the information exchange problem, and is oriented to express resources and relations in a standard way (in agreement to the RDF philosophy); GDM [Hid02] defines instances and schema graphs with features similar to RDF (e.g. domain and range of relations, typeOf properties). Models like G-BASE [Kun87], Gram [AS92], GraphDB [GÖ94] and GRAS [KSW96] propose explicit graph data models.[5]. Besides these models based on graphs, there are other approaches which use as formalization generalizations of the notion of graph, such as hypergraphs (e.g. see GROOVI [LP91], the hypernode model [LP90,PL94]) and hygraphs (e.g. see Hy+ [CM93]). Note that strictly speaking, RDF graphs are ordered hypergraphs [HG04]. As was mentioned above, an important component of these data models is the query

[5] Note that a direct applicability of a graph model to RDF is not possible due to the particular RDF graph property where resources possibly can occur as edge labels as as well as node labels. To solve this problem an intermediate model (e.g bipartite graphs [HG04]) can be defined.

language, which differs widely from one to another, and whose design is very challenging. We survey them in the next section.

3.2 Graph query languages

There are several proposals of query languages for models that represent information with a explicit or implicit graph structure. In this context, from now on we assume that a graph database has n nodes and e edges.

Cruz et al. [CMW87] propose the graphical query language G for querying data represented as a labeled graph. It introduces the concept of graphical query, which is based on a pattern graph that use regular expressions to represent recursive queries. G evolved into a more powerful language called G+ [BHK$^+$03] where a query graph is the basic building block. Query graph nodes may be labeled with variables and edges labeled with regular expressions. A simple query has two elements, a query graph that specifies the class of patterns to search and a summary graph that represent how to restructure the answer obtained by the query graph.

GraphLog [CM89] is a query language for hypertext. It presents a extension of G+ by adding negation and unifying the concept of a query graph. A query is now only one graph pattern containing one distinguished edge (which corresponds to the restructured edge of the summary graph in G+). The effect of the query is to find all instances of the pattern that occur in the database graph and for each one of them define a virtual link represented by the distinguished edge.

Gram [AS92] presents a query algebra where regular expressions over data types are used to select walks (paths) in a graph. It uses a data model where walks are the basic objects. A walk expression is a regular expression without union, whose language contains only alternating sequences of node and edge types, starting and ending with a node type. The query language is based on a hyperwalk algebra with operations closed under the set of hyperwalks.

Gemis and Paredaens [GP93] present PaMal, a graphical model for describing schemes and instances of object-databases and a graphical data manipulation language based on pattern matching.

Güting [GÖ94] proposed an object-oriented data model and query language for graph databases called GraphDB. A database in GraphDB is a collection of object classes divided in: simple classes (simple objects that represent nodes), link classes (links between nodes that represent edges) and path classes (representing several paths in the database). A query consists of several steps. Each step computes operations that specify argument subgraphs in the form of regular expressions over link class names that extend or restrict dynamically the database graph.

Lorel [AQM$^+$97] is a query language for semistructured data designed for the Object Exchange Model (OEM) [PGMW95]. Lorel is a extension of OQL [ASL89], extending its characteristics to handling semistructured data.

Oriented to search the Web, Flesca and Grego [FG99] extend regular expressions and regular grammars by introducing partial orders on strings and production rules, respectively to support graph queries. They show to how use partially

ordered languages to define path queries to search databases and present results on their computational complexity. In addition, a query language based on the previous ideas was proposed in [FG00].

4 Current RDF query languages and their graph support

In this section we present the current state of RDF query languages. First, for the sake of completeness, we briefly review current proposals, and the activity of the W3C Working Group on Data Access. Then we test the most relevant of these languages against simple queries which have graph-like flavor to show the degree of support for querying in the graph databases style.

4.1 Overview of RDF query languages

RDF Query Languages. Several languages for querying RDF data have been proposed and implemented, some in the lines of traditional database query languages (e.g. SQL, OQL), others based on logic and rule languages. Good surveys are [MKA+02,HBEV04].

RQL [KAC+02] is a typed language for querying RDF repositories. It is defined by a set of basic queries and iterators, which can be used to build new ones through functional decomposition, with a OQL style. *SquishQL*[6] is a SQL-style query language that permits simple graph navigation in RDF sources. It uses patterns and filter expressions to construct queries and use a subgraph matching mechanism to execute the query. *RDQL* [Sea04] is an implementation of *SquishQL* for *Jena* [WSKR03] models. *RDFQL*[7] is a statement-based query language with a SQL-style to perform queries, inference operations, and construction of views on RDF structured data. *RDFQL* is capable to infer new statements from existing ones by using user-defined inference rules. *TRIPLE* [SD02] is a language based on *F-logic* [KL89], allowing rule definition, inference and transformation of RDF models under several semantics. The core language is based on Horn logic syntactically extended to support RDF primitives (namespaces, resources, and statements). *Notation 3* (N3) [BL01] provides a text-based syntax for RDF, allowing to define rules (stored with the data) that can be used for the purpose of querying. *Versa*[8] is a graph-based language with some support for rules. Its main features include traversal arcs, processing of node contents and general expression evaluation. *SeRQL* is the RDF query language developed as part of the *Sesame* system[9]. It combines characteristics of languages like *RQL, RDQL, N-Triple, N3* plus some new features. *SeRQL* presents support for graph transformation, RDF schema, expressive path expression syntax, optional path

[6] `http://ilrt.org/discovery/2001/02/squish/`
[7] `http://www.intellidimension.com/default.rsp?topic=/pages/rdfgateway/reference/db/default.rsp`
[8] `http://4suite.org/`
[9] `http://www.openrdf.org/`

matching, and composition of queries. *RXPath*[10] is a query language based on XPath, that allows addressing parts of a RDF model with a syntax identical to XPath. *Turtle*[11] is a text syntax for RDF that extends carefully the test case format of *N-Triples* [Bec01], taking the most useful and appropriate features from *Notation 3*.

RDF Data Access Working Group (DAWG) – SPARQL. W3C members that conform the RDF DAWG presented a Working Draft in October 2004, which specifies a set of use cases, requirements, and objectives for a RDF query language and data access protocol [Cla04]. The use cases are directed to application areas of RDF, like personal and business information management, resource publishing, transportation, software development, social networks, and semantic web. The technical requirements can be grouped into *data model support* (limited datatype support, extensible value testing), *evaluation of graphs* (RDF graph pattern matching conjunction/disjunction, variable binding results), *construction of output* (subgraph results, local queries), and *output management* (optional match, result limits, streaming results). The design objectives are human-friendly syntax, data integration and aggregation, test of non-existence of triples, bandwidth-efficient protocol for a collection of result, literal search, express yes-no queries directly and addressable query results.

SPARQL [PS04] is a RDF query language designed to meet the requirements and design objectives mentioned previously. It defines a query language with a SQL-like style, where a simple query is based on query patterns, and query processing consists of binding of variables to generate pattern solutions (graph pattern matching). *SPARQL* supports multiple matches (multiple selected variables), constraining values in the form of Boolean-valued expressions, optional matching that permits blank results, nested patterns (e.g. simple conjunction), execution against real or virtual graphs, results construction, and standard operations (which are a subset of the operations defined in XQuery and XPath). *SPARQL* is still a work in progress.

4.2 Graph Properties in Current RDF Query Languages

To illustrate the problems of current RDF query languages in querying graph-like properties, we chose seven of them, six already considered in the recent survey [HBEV04], and RxPath, and test them against a subset of the well-known Museum example (see Figure 1), which –despite its small size– proved well able to illustrate our point. To simplify, we did not consider the issue of support for subclass or subproperty semantics.

The results are as follows. An RDF graph can be considered a directed graph. This direction produces problems when retrieving neighborhoods for languages that do not have a union operator (see queries 1 and 2 below). Some query results violate the query language property of closure [HBEV04] by returning

[10] `http://rx4rdf.liminalzone.org/`
[11] `http://www.ilrt.bris.ac.uk/discovery/2004/01/turtle/`

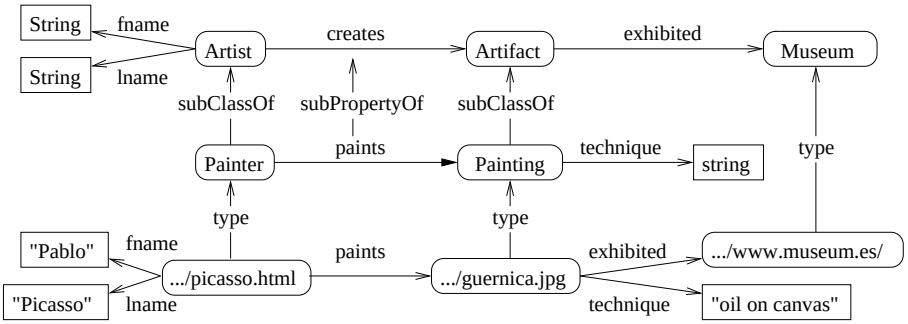

Fig. 1: The subset of the Museum database [KAC+02] used for our tests

results which are not in RDF format. There are two main problems concerning paths: (a) most languages support only querying for patterns of paths which are limited in length and form (the issue of edge direction blows up the size of the query exponentially, see below query 4 and 5); (b) RxPath is able to retrieve only paths starting from a fixed node and with some other restrictions. Aggregated functions like COUNT, MIN, MAX applied to paths could be used to answer queries as for the degree of a node, the distance between nodes, and the diameter of a graph. None of these functions is systematically supported, even though, for example, the original version of RQL has a COUNT function on the number of triples. Next we present the details of each query. Table 2 summarizes the results (implementation details can be found in [AGH04]).

1. *Adjacent nodes:* "All resources adjacent to the resource `Guernica`". Expected result: `Painting`, `www.museum.es`, `"oil on canvas"`, `picasso.html`. Not all languages support this feature. The problem is that this query can only be expressed as a union of two queries: one for outgoing edges from `Guernica`, another for ingoing edges. Some languages do not support the union operator.
2. *Adjacent edges:* "All predicates of statements involving `Guernica`". Expected result: `technique`, `exhibited`, `type`, `paints`. The problems faced are similar to the previous case. Note that here we probably would like to differentiate schema predicates from data predicates.
3. *Degree of a node:* "Number of predicates involving `Guernica`". Expected result: 4. Same problems as above plus the fact that most languages do not support aggregation at this level. SeRQL for example returns the number, but not as part of the answer.
4. *Path:* "Find paths between `picasso.html` and `www.museum.es`". Expected results: There are several, for example, `picasso.html`–`paints`–`guernica.jpg`–`exhibited`–`www.museum.es`, and `picasso.html`–`type`–`Painter`–`paints`–`Painting`–`subClassOf`–`Artifact`–`exhibited`–`Museum`–`type`–`www.museum.es`.

PROPERTY	RQL	SeRQL	RDQL	Triple	N3	Versa	RxPath
Adjacent nodes	±	±	±	±	±	±	×
Adjacent edges	±	±	±	±	×	×	×
Degree of a node	±	×	×	×	×	×	×
Path	×	×	×	×	×	×	±
Fixed-length Path	±	±	±	±	±	×	±
Distance between two nodes	×	×	×	×	×	×	×
Diameter	×	×	×	×	×	×	×

Table 2: Support of some current RDF query languages for some example queries ("±" indicates partial support and "×" no support).

None of the languages studied support arbitrary paths like the ones needed for this case. Note that it must be considered whether paths via the schema are regarded as relevant.

5. *Fixed-length paths.* "Find all paths of length 2 between `Pablo` and `guernica.jpg`." Expected result: `Pablo`–`fname`–`picasso.html`–`paints`–`guernica.jpg`. Supported partially by several languages, using a union of all possibles patterns of paths (direction of edges) of length 2 between the initial and final resources. Note that for paths of length n there are at least 2^n such path patterns.
6. *Distance between two resources:* (length of shortest path) "How far is `picasso.html` from `www.museum.es`?" Expected result: 2. Not supported by any language.
7. *Diameter of a graph:* "Diameter of the museum graph". Expected result: 5. Not supported. It is based on distance and paths.

5 Graph primitives for RDF query languages

In this section we present desirable graph primitives of a query language for the RDF data model, based on the experience of the graph database query languages discussed in previous sections. We stress the graph-like features that in our opinion are missing in today's RDF query languages.

Before discussing the primitives in detail, let us enumerate desirables features for a RDF query language. They are very much inspired by a similar wish-list stated by Abiteboul [Abi97] for semi-structured data. They are:

- Standard database-style query primitives.
- Navigation in the style of semi-structured data or Web-style browsing.
- Searching for patterns in an information-retrieval style.
- Temporal queries, including versioning.
- Querying both the data and the schema in the same query.
- Incorporating transparently the lightweight inferencing of RDF Schema and relevant polynomial-time extensions.
- Sound theoretical foundation.

PROPERTY	G	G+	GraphLog	Gram	GraphDB	Lorel	F-G
Adjacent nodes	$\pm$	$\surd$	$\surd$	$\surd$	$\pm$	$\surd$	$\pm$
Adjacent edges	$\pm$	$\surd$	$\surd$	$\surd$	$\pm$	$\surd$	$\pm$
Degree of a node	$\times$	$\surd$	$\surd$	$\times$	?	$\times$	$\times$
Path	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Fixed-length Path	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Distance between two nodes	$\times$	$\surd$	$\surd$	$\times$	?	$\times$	$\times$
Diameter	$\times$	$\surd$	$\surd$	$\times$	?	$\times$	$\times$

Table 3: Support of some graph database query languages for the example queries of Table 2 ($\surd$ indicates support, "$\pm$" partial support, "$\times$" no support, and ? indicates there is no information available).

The following groups of primitives comprise features revised in the survey and the ones presented in the test examples we think should be supported by an RDF query language. In each case we survey the support that graph database languages gives them. See also Table 3 for examples of support for some queries by these languages (compare to Table 2).

Paths and connectedness. One of the most fundamental graph problems is to compute reachability information (use case 2.5 in DAWG Draft [Cla04]). In fact, many of the recursive queries that arise in practice in relational databases and, more generally in data with graph structure, are in practice graph traversals characterized by path problems. The importance of such queries is studied in several works [AJ94,AJ89,AJ88,RLMB98]. One of the challenges to incorporate such notion into a query language is its computational complexity. Finding simple paths with desired properties in direct graphs is very difficult, and essentially every nontrivial property gives rise to an NP-complete problem [SWG02]. Yannakakis [Yan90] surveyed a set of paths problems relevant to the database area including computing transitive closures, recursive queries and the complexity of path searching. Extension of query languages for solve graph traversal problems are surveyed in [MS90].

In what follows, we describe the support that the query languages of the database models described in Section 3.1 give to path problems.

A initial implementation of G translate the graphical queries into C-Prolog programs. Simple paths are traversed using certain non-Horn clause constructs available in Prolog. Although, it does not support cycles or finding the shortest path, it was a good approximation to a graph query language.

The evaluation of path queries in G+ is a two-stage process consisting of a depth-first search of the graph database and use of a nondeterministic finite state automaton to control the search. In addition path queries are a subset of the class of linear chain queries and hence can be evaluated rapidly in parallel. The evaluation algorithm can be shown to compute the identity query in O(e) time and the transitive closure in O(ne) time. G+ was implemented in the

HyperG system providing primitive operators like depth-first search, shortest path, transitive closure and connected components.

Motivated by the implementation of G+, Mendelzon and Wood [MW89] studied the problem of finding all pair of nodes connected by a simple path such that the concatenation of the labels along the path satisfies a regular expression. Although the regular simple path problem is in general NP-complete, the paper presents an algorithm that runs in polynomial time in the size of the graph when some conditions fulfilled: the graph is acyclic, the regular expression is restricted (according to the definition in the paper), or the graph complies with a cycle constraint compatible with the regular expression. The evaluation algorithm uses a deterministic finite automaton to traverse paths in the graph. They also prove the intractability of certain types of simple paths in a particular class of direct graphs and characterize a class of queries about regular simple paths which can be evaluated in polynomial time. The analysis and implementation in this paper, assume that the graph can be entirely stored in main memory.

The expressive power of GraphLog is characterized by establishing the equivalence between GraphLog, stratified linear Datalog (a language of function-free Horn clauses), non deterministic logarithmic space, and transitive closure. The queries expressible in the language are exactly those that can be computed in space logarithm in the size of the database.

To implement graph operations in GraphDB, efficient graph algorithms are used. Shortest path and cycle both were implemented using the A* algorithm. Moreover, nodes, paths and subgraphs are indexed using path classes and index structures like B-Tree and LSD-Tree.

Lorel presents a SQL-style query language that support two types of path expressions, simple path expressions, which allow to obtain the set of objects reachable by following a sequence of labels starting from a named object in the OEM graph and a more powerful syntax for path expressions, called general path expressions based on wildcards and regular expressions. To outperform query execution, the Lore DBMS [MAG$^+$97] implements the query language Lore and uses two kinds of indexes, a link (edge) index called Lindex, and a value index called Vindex. A Lindex takes an object identifier and a label, and returns the object identifiers of all parents via the specified label. A Vindex takes a label, operator, and a value, and returns all atomic objects having an incoming edge with the specific label and a value satisfying the specific operator and value. Vindexes and Lindexes are implemented using B+ trees and linear hashing respectively.

In graph databases where the number of nodes is very large (e.g. the Web) it is useful to subdivide the domain of evaluation by selecting subsets of the domain on the base of some criteria. With this objective, Flesca and Greco [FG99] introduce partially ordered regular languages based on some order on the nodes. Such languages are an extension of regular languages where strings are partially ordered, for example, two strings s_1 and s_2, such that $s_1 > s_2$, denote two paths in the graph with the constraint that the path s_1 should be preferred to the path s_2. In later work [FG00], they present an algebra for partially ordered relations,

an algorithm for the computation of path queries and show that computing an instance of a graph query can be done in polynomial time. Also, they present a SQL-like language that consider general paths and extended regular expressions, and show how extended regular expressions can be used to search the Web. With similar motivations, and in the context of RDF, Anyanwu and Sheth [AS03] introduced a path operator ρ to address relevant relationships between entities called semantic associations. Semantic associations are represented in a RDF graph as sequences (i.e. edges, paths) between entities or more complex structures of sequences, and a notion of similarity between them is defined.

Pattern matching consists in determining if there exists a mapping (or isomorphism) between a graph pattern and a subgraph of a database graph (use cases 2.1, 2.12 and 2.13 in DAWG Draft [Cla04]). Pattern matching deal with two problems, the graph isomorphism problem that has a unknown computational complexity, and the subgraph isomorphism problem which is NP-complete. Pattern matching has attracted a great deal of attention specially on data mining (see [WM03] for a survey), update [GP93,HP93], querying [CMW87,CMW89,CM89] and visualization [AGP+92]. Sasha *et al.* [SWG02] present a survey of pattern-matching based algorithms for fast searching in trees and graphs.

PaMal use graph patterns to describes the part of the database instance that are affected by a operation (addition and deletion of nodes and edges). In the case of GraphDB, the subgraph problem is solved moving the conditions into subsequent graph operations or other database access.

Aggregate functions are operations non related to the data model that permit to summarize or operate on the query results (use cases 2.3, 2.4, 2.6, 2.8, 2.10, 2.11, 2.14 and 2.15 in DAWG Draft [Cla04]). Such functions are oriented to deal directly with the structure of the underlying graph, such as the degree of a node, the diameter of the graph (or a set of nodes), the distance between nodes, etc.

With the purpose of performing computations on retrieved subgraphs product of a query operation, G+ defines two types of summary operators: path operators which summarize on the values of the attributes along paths and set operators which summarize on the values of the attributes on a set of paths. The set of such operators include sum, products, maximum and count.

GraphLog becomes more expressive that relational algebra and calculus with aggregates, adding aggregate operators (e.g. MAX, SUM, etc.) and path summarization. The implementation of GraphLog use algorithms discussed in [MW89].

Gram, consistent with its SQL-like syntax, defines two types of algebraic operations: unary (projection, selection, renaming) and binary (join, concatenation, set operations) which are closed under the set of hyperwalks. PaMal provides a reduce-operation to work with a special group of instances called reduced instances and programming constructs (loop, procedure and program). Finally, GraphDB query language support further operations, e.g. for sorting, grouping, and aggregate functions (e.g. Sum).

Neighborhoods. The notion of neighborhood is relevant for information having a graph-like nature (use case 2.7 in DAWG Draft [Cla04]). In these models, information (represented by nodes) closed (in the graph) is usually semantically related . The primary notion is *adjacency.* Both node and edge adjacency in RDF are important in various contexts. A more advanced notion of adjacency, like *the k-neighborhood of a node*, is necessary in several contexts. The need of 1-neighborhood retrieval in a RDF Graph is argued in [Say04] and [GMM03]. In the RDF context, inference of new triples is relevant in Vertex and Edge adjacency queries. To the best of our knowledge, the notion of neighborhood as primitive for query languages has not been studied systematically in the database literature.

6 Conclusions

We considered RDF from the perspective of graph database modeling. We compared it with other database models. We surveyed graph database models and query languages in order to argue the convenience that the RDF community incorporate database experience and technologies into further development of the RFD model and query language design. In concrete, we propose that RDF query language should incorporate graph database query language primitives. Further work includes developing use cases, formalizing requirements and building benchmarks for queries using the graph-like structure of the model.

References

[Abi97] Serge Abiteboul. Querying Semi-Structured Data. In *Proceedings of the 6th International Conference on Database Theory (ICDT'97)*, volume 1186 of *Lecture Notes in Computer Science*, pages 1–18. Springer, January 8-10 1997.

[ADL+04] Sinuhe Arroyo, Ying Ding, Ruben Lara, Michael Stollberg, and Dieter Fensel. Semantic Web Languages. Strengths and Weakness. In *International Conference in Applied computing (IADIS04)*, March 23-26 2004.

[AGH04] Renzo Angles, Claudio Gutierrez, and Jonathan Hayes. RDF Query Languages Need Support for Graph Properties. Technical Report TR/DCC-2004-3, Department of Computer Science, University of Chile, June 2004.

[AGP+92] Marc Andries, Marc Gemis, Jan Paredaens, Inge Thyssens, and Jan Van den Bussche. Concepts for Graph-Oriented Object Manipulation. In *3rd International Conference on Extending Database Technology (EDBT'92)*, volume 580 of *Lecture Notes in Computer Science*, pages 21–38. Springer, March 23-27 1992.

[AJ88] Rakesh Agrawal and H. V. Jagadish. Efficient Search in Very Large Databases. In *Proceedings of the 14th International Conference on Very Large Data Bases*, pages 407–418. Morgan Kaufmann, August 29 - September 1 1988.

[AJ89] Rakesh Agrawal and H. V. Jagadish. Materialization and Incremental Update of Path Information. In *Proceedings of the Fifth International*

Conference on Data Engineering, pages 374–383. IEEE Computer Society, February 6-10 1989.

[AJ94] R. Agrawal and H. V. Jagadish. Algorithms for Searching Massive Graphs. *IEEE Transactions on Knowledge and Data Engineering*, 6(2):225–238, 1994.

[AQM+97] Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, and Janet L. Wiener. The Lorel query language for semistructured data. *International Journal on Digital Libraries*, 1(1):68–88, 1997.

[AS92] Bernd Amann and Michel Scholl. Gram: A Graph Data Model and Query Language. In *European Conference on Hypertext Technology (ECHT '92)*, pages 201–211. ACM, November 30 - December 4 1992.

[AS03] Kemafor Anyanwu and Amit Sheth. The ρ-operator: Enabling Querying for Semantic Associations on the Semantic Web. In *The Twelfth International World Wide Web Conference (WWW2003)*, 20-24 May 2003.

[ASL89] A. M. Alashqur, S. Y. W. Su, and H. Lam. OQL: a query language for manipulating object-oriented databases. In *Proceedings of the fifteenth international conference on Very Large Data Bases*, pages 433–442. Morgan Kaufmann Publishers Inc., 1989.

[Bec01] Dave Beckett. N-Triples - W3C RDF Core WG Internal Working Draft. http://www.w3.org/2001/sw/RDFCore/ntriples/, 2001.

[Bee88] Catriel Beeri. Data Models and Languages for Databases. In *Proceedings of the 2nd International Conference on Database Theory (ICDT'88)*, volume 326 of *Lecture Notes in Computer Science*, pages 19–40. Springer, August 31 - September 2 1988.

[BHK+03] Andrey Balmin, Vagelis Hristidis, Nick Koudas, Yannis Papakonstantinou, Divesh Srivastava, and Tianqiu Wang. A System for Keyword Proximity Search on XML Databases. In *Proceedings of 29th International Conference on Very Large Data Bases, VLDB 2003*, pages 1069–1072. Morgan Kaufmann, September 9-12 2003.

[BL01] Tim Berners-Lee. Notation 3 - An RDF language for the Semantic Web. http://www.w3.org/DesignIssues/Notation3, 2001.

[BPSM] Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0, W3C Recommendation 10 February 1998. http://www.w3.org/TR/1998/REC-xml-19980210.

[Bun97] Peter Buneman. Semistructured Data. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 117–121. ACM Press, May 12-14 1997.

[BV04] Nick Bassiliades and Ioannis P. Vlahavas. R-device: A deductive rdf rule language. In *Proceedings of the Third International Workshop, Rules and Rule Markup Languages for the Semantic Web (RuleML)*, pages 65–80, 2004.

[Che76] Peter Pin-Shan Chen. The entity-relationship model-toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.

[Cla04] Kendall Grant Clark. RDF Data Access Use Cases and Requirements, W3C Working Draft. http://www.w3.org/TR/rdf-dawg-uc/, October 12 2004.

[CM89] M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in graphlog. In *Proceedings of the second annual ACM conference on Hypertext*, pages 269–292. ACM Press, 1989.

[CM93] Mariano P. Consens and Alberto O. Mendelzon. Hy+: A Hygraph-based Query and Visualization System. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 511–516. ACM Press, May 26-28 1993.

[CMW87] Isabel F. Cruz, Alberto O. Mendelzon, and Peter T. Wood. A Graphical Query Language Supporting Recursion. In *Proceedings of the Association for Computing Machinery Special Interest Group on Management of Data*, pages 323–330. ACM Press, May 27-29 1987.

[CMW89] I. F. Cruz, A. O. Mendelzon, and P. T. Wood. G+: Recursive Queries without Recursion. In *Proc. of the Second International Conference on Expert Database Systems*, pages 645–666. Addison-Wesley, April 25-27 1989.

[Cod80] E. F. Codd. Data models in database management. In *Proceedings of the 1980 workshop on Data abstraction, databases and conceptual modeling*, pages 112–114. ACM Press, 1980.

[Cod83] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 26(1):64–69, 1983.

[FG99] Sergio Flesca and Sergio Greco. Partially Ordered Regular Languages for Graph Queries. In *Proceedings of the 26th International Colloquium on Automata, Languages and Programming, ICALP'99*, volume 1644 of *Lecture Notes in Computer Science*. Springer, July 11-15 1999.

[FG00] Sergio Flesca and Sergio Greco. Querying Graph Databases. In *Advances in Database Technology - EDBT 2000, Proceedings of the 7th International Conference on Extending Database Technology*, volume 1777 of *Lecture Notes in Computer Science*, pages 510–524. Springer, March 27-31 2000.

[GÖ94] Ralf Hartmut Güting. GraphDB: Modeling and Querying Graphs in Databases. In *Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 297–308. Morgan Kaufmann, September 12-15 1994.

[GHO04] Claudio Gutierrez, Carlos Hurtado, and O.Mendelzon. Foundations of Semantic Web Databases. In *Proceedings of the 33rd ACM Symposium on Principles of Database Systems*, June 14-16 2004.

[GMM03] R. Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the twelfth international conference on World Wide Web*, pages 700–709. ACM Press, 2003.

[GP93] Marc Gemis and Jan Paredaens. An Object-Oriented Pattern Matching Language. In *Proceedings of the First JSSST International Symposium on Object Technologies for Advanced Software*, pages 339–355. Springer-Verlag, 1993.

[GPdBG90] Marc Gyssens, Jan Paredaens, Jan Van den Bussche, and Dirk Van Gucht. A Graph-Oriented Object Database Model. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 417–424, April 2-4 1990.

[GR02] Yolanda Gil and Varun Ratnakar. A Comparison of (Semantic) Markup Languages. In *Proceedings of the 15th International FLAIRS Conference*, May 14-16 2002.

[Hay04] Patrick Hayes. RDF Semantics. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/, Feb 2004.

[HBEV04] Peter Haase, Jeen Broekstra, Andreas Eberhart, and Raphael Volz. A Comparison of RDF Query Languages. In *Third International Semantic Web Conference (ISWC2004)*, November 2004.

[HG04] Jonathan Hayes and Claudio Gutierrez. Bipartite Graphs as Intermediate Model for RDF. In *Third International Semantic Web Conference (ISWC2004)*, November 2004.

[Hid02] Jan Hidders. Typing Graph-Manipulation Operations. In *Proceedings of the 9th International Conference on Database Theory*, pages 394–409. Springer-Verlag, 2002.

[HP93] Jan Hidders and Jan Paredaens. GOAL, A Graph-Based Object and Association Language. *Advances in Database Systems: Implementations and Applications, CISM*, pages 247–265, September 13-24 1993.

[HT02] Ian Horrocks and Sergio Tessaris. Querying the Semantic Web: a Formal Approach. In *Proc. of the 13th Int. Semantic Web Conf. (ISWC 2002)*, number 2342 in Lecture Notes in Computer Science, pages 177–191. Springer-Verlag, 2002.

[KAC+02] Gregory Karvounarakis, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, and Michel Scholl. RQL: a declarative query language for RDF. In *Proceedings of the eleventh international conference on World Wide Web*, pages 592–603. ACM Press, 2002.

[KC04] Graham Klyne and Jeremy Carroll. Resource Description Framework (RDF) Concepts and Abstract Syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, Feb 2004.

[Kim90] W. Kim. Object-Oriented Databases: Definition and Research Directions. *IEEE Transactions on Knowledge and Data Engineering*, 2(3):327–341, 1990.

[KKT76] Larry Kerschberg, Anthony C. Klug, and Dennis Tsichritzis. A Taxonomy of Data Models. In *Systems for Large Data Bases*, pages 43–64. North Holland and IFIP, September 8-10 1976.

[KL89] Michael Kifer and Georg Lausen. F-Logic: A Higher-Order language for Reasoning about Objects, Inheritance, and Scheme. In *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, pages 134–146. ACM Press, May 31 - June 29 1989.

[KSW96] Norbert Kiesel, Andy Schurr, and Bernhard Westfechtel. GRAS: A Graph-Oriented Software Engineering Database System. In *IPSEN Book*, pages 397–425, 1996.

[Kun87] H. S. Kunii. DBMS with graph data model for knowledge handling. In *Proceedings of the 1987 Fall Joint Computer Conference on Exploring technology: today and tomorrow*, pages 138–142. IEEE Computer Society Press, 1987.

[KV84] Gabriel M. Kuper and Moshe Y. Vardi. A New Approach to Database Logic. In *Proceedings of the 3th ACM SIGACT-SIGMOD Symposium on Principles of Database Systems*, pages 86–96. ACM, April 2-4 1984.

[KV93] Gabriel M. Kuper and Moshe Y. Vardi. The Logical Data Model. *ACM Transactions on Database Systems*, 18(3):379–413, 1993.

[LP90] M. Levene and A. Poulovassilis. The hypernode model and its associated query language. In *Proceedings of the fifth Jerusalem conference on Information technology*, pages 520–530. IEEE Computer Society Press, 1990.

[LP91] M. Levene and A. Poulovanssilis. An object-oriented data model formalised through hypergraphs. *Data Knowledge and Engineering*, 6(3):205–224, 1991.

[LRV88] Christophe Lécluse, Philippe Richard, and Fernando Vélez. O2, an Object-Oriented Data Model. In *Proceedings of the 1988 ACM SIGMOD Inter-*

national Conference on Management of Data, pages 424–433. ACM Press, June 1-3 1988.

[LS99] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, February 1999.

[MAG+97] Jason McHugh, Serge Abiteboul, Roy Goldman, Dallan Quass, and Jennifer Widom. Lore: A Database Management System for Semistructured Data. *SIGMOD Record*, 26(3):54–66, 1997.

[MKA+02] Aimilia Magkanaraki, Grigoris Karvounarakis, Ta Tuan Anh, Vassilis Christophides, and Dimitris Plexousakis. Ontology Storage and Querying. Technical Report 308, Institute of Computer Science of the Foundation for Research and Technology - Hellas (ICS-FORTH), April 2002.

[MS90] M. V. Mannino and L. D. Shapiro. Extensions to Query Languages for Graph Traversal Problems. *IEEE Transactions on Knowledge and Data Engineering*, 2(3):353–363, 1990.

[MW89] A. O. Mendelzon and P. T. Wood. Finding regular simple paths in graph databases. In *Proceedings of the fifteenth international conference on Very large data bases*, pages 185–193. Morgan Kaufmann Publishers Inc., 1989.

[Nav92] Shamkant B. Navathe. Evolution of data modeling for databases. *Communications of the ACM*, 35(9):112–123, 1992.

[Olk03] Frank Olken. Tutorial on Graph Data Management for Biology. *IEEE Computer Society Bioinformatics Conference (CSB)*, Aug 2003.

[PGMW95] Yannis Papakonstantinou, Hector Garcia-Molina, and Jennifer Widom. Object Exchange across Heterogeneous Information Source. In *11th Conference on Data Engineering*, pages 251–260, Taipei, Taiwan, 1995. IEEE Computer Society.

[PL94] Alexandra Poulovassilis and Mark Levene. A nested-graph model for the representation and manipulation of complex objects. *ACM Transactions on Information Systems*, 12(1):35–68, 1994.

[PM88] Joan Peckham and Fred J. Maryanski. Semantic Data Models. *ACM Computing Surveys*, 20(3):153–189, 1988.

[PS04] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/, October 12 2004.

[RLMB98] R.V.Guha, Ora Lassila, Eric Miller, and Dan Brickley. Enabling Inferencing. *The Query Languages Workshop, QL'98*, December 1998.

[SAMA+05] Amit Sheth, Boanerges Aleman-Meza, I. Budak Arpinar, Christian Halaschek-Wiener, Cartic Ramakrishnan, Clemens Bertram, Yashodhan Warke, David Avant, F. Sena Arpinar, Kemafor Anyanwu, and Krys Kochut. Semantic association identification and knowledge discovery for national security applications. *Journal of Database Management*, 16(1):33–53, Jan-March 2005.

[Say04] Craig Sayers. Node-centric RDF Graph Visualization. Technical Report HPL-2004-60, Mobile and Media Systems Laboratory HP Laboratories Palo Alto, April 2004.

[SD02] Michael Sintek and Stefan Decker. TRIPLE - A Query, Inference, and Transformation Language for the Semantic Web. *International Semantic Web Conference (ISWC)*, June 2002.

[Sea04] Andy Seaborne. RDQL - A query Language for RDF, W3C Member Submission 9 January 2004. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/, Jan 2004.

[Shi81] David W. Shipman. The functional data model and the data languages DAPLEX. *ACM Transactions on Database Systems (TODS)*, 6(1):140–173, 1981.

[SKS96] Avi Silberschatz, Henry F. Korth, and S. Sudarshan. Data models. *ACM Computing Surveys*, 28(1):105–108, 1996.

[Str88] Bjarne Stroustrup. What Is Object-Oriented Programming? *IEEE Softw.*, 5(3):10–20, 1988.

[SWG02] Dennis Shasha, Jason T. L. Wang, and Rosalba Giugno. Algorithmics and applications of tree and graph searching. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 39–52. ACM Press, 2002.

[TF76] Robert W. Taylor and Randall L. Frank. CODASYL Data-Base Management Systems. *ACM Comput. Surv.*, 8(1):67–103, 1976.

[TL76] D. C. Tsichritzis and F. H. Lochovsky. Hierarchical Data-Base Management: A Survey. *ACM Comput. Surv.*, 8(1):105–123, 1976.

[WM03] Takashi Washio and Hiroshi Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1):59–68, 2003.

[WSKR03] Kevin Wilkinson, Craig Sayers, Harumi Kuno, and Dave Reynolds. Efficient RDF Storage and Retrieval in Jena2. In *Proceedings of SWDB'03*, 2003.

[Yan90] Mihalis Yannakakis. Graph-theoretic methods in database theory. In *Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 230–242. ACM Press, 1990.