Links and Paths through Life Sciences Data Sources

Zoé Lacroix
Arizona State University
zoe.lacroix@asu.edu

Felix Naumann
Humboldt-University of Berlin
naumann@informatik.hu-berlin.de

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Hyma Murthy
University of Maryland
hmurthy@umiacs.umd.edu

Abstract

An abundance of biological data sources contain data on classes of scientific entities, such as genes and sequences. Logical relationships between scientific objects are implemented as URLs and foreign IDs. Query processing typically involves traversing links and paths (concatenation of links) through these sources. We model the data objects in these sources and the links between objects as an object graph. We identify a set of interesting properties for links and paths, such as outdegree, image of a link, cardinality of data objects and links, the number of distinct objects reached by some links, etc. Analogous to database cost models, we use statistics from the object graph to develop a framework to estimate the result size for a query on the object graph. Analogous to training and testing, we use sampled data from queries to estimate the result size. We validate our models using data sampled from four NIH/NCBI data sources. Our research provides a foundation for querying and exploring data sources.

1 Querying Interlinked Sources

An abundance of biological data sources contain data about scientific entities, such as genes and sequences. Each source may have data on one or more logical classes. Logical relationships between scientific objects are implemented as *source links* between data sources. Together, they form a graph – the *source graph*. Each source link represents a collection of object links, each going from a data object in one source to another object, in the same or a different source. An *object graph* is formed of the data objects and links. Formal definitions are in Sec. 2.

Scientists are interested in exploring relationships between scientific objects, e.g., genes and citations. Consider the query "*Return all citations of* PubMed *that are linked to an* Omim *entry that is related to some disease or condition.*" To answer such queries, biologists and query engines alike must fully traverse links and paths (informally concatenations of links) through these sources given some start object in Omim. Fig. 1 illustrates the source graph for four data sources at the National Center for Biotechnology Information (NCBI). A scientist may choose the Omim source, which contains information related to human genetic diseases, as a starting point for her exploration and wish to eventually retrieve citations from the PubMed source. Starting with a keyword search on a certain disease, she can explore direct links between genes in Omim and citations in PubMed. She can also traverse paths that are implemented using additional intermediate sources to learn about this relationship.

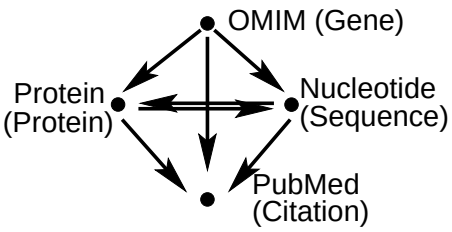

Figure 1: Source graph for NCBI data sources (and corresponding scientific entities)

In all, there are five paths (without loops) starting from Omim and terminating in PubMed. These paths are shown in Fig. 2.

(P1) Omim → PubMed
(P2) Omim → Nucleotide → PubMed
(P3) Omim → Protein → PubMed
(P4) Omim → Nucleotide → Protein → PubMed
(P5) Omim → Protein → Nucleotide → PubMed

Figure 2: All five paths from Omim to PubMed through the source graph of Fig. 1

The choice of paths has an impact on the result. For example, traversing a path via the Protein source might yield less and different citations compared to a path via the Nucleotide source. This depends on the intermediate sources and corresponding entity classes that are traversed in a path, the contents of each source, the contents of each source link, etc.

These properties of paths and their effects are of interest from a number of perspectives: From a *query evaluation* viewpoint, one can estimate the cost and benefit of evaluating a query given some specific sources and paths. In traditional database query optimization, cost models are typically used to estimate the execution time, using statistics such as table cardinalities, selectivity factors of certain operations, derived cardinalities of intermediate results, etc. In analogy, we use source cardinalities, link cardinalities etc., to estimate the size of the intermediate and the final result. While the query execution time is important to researchers, especially when sources are remote, they are equally interested in the quality of the result. Quality covers multiple aspects such as completeness of the source contents and link contents, reputation of the data providers, etc. This paper provides a basis for the comparison of properties of different paths through interlinked data sources and thus a means for optimization that recognizes the importance of both execution time as well as quality of the results.

A second perspective that can profit from this work is that of *data curation*. Administrators of cross-linked data sources are interested in providing not only correct data, but also complete and consistent links to related data items in other data sources. For example, GeneReviews is an alternate source of genetic information on diseases and it too has links to citations in PubMed. The results presented in this paper can help curators gain insight into the link structure of their data. For example, we can compare the results of traversing the source link from Omim to PubMed and compare the results with traversing the source link from GeneReviews to PubMed. Based on our analysis, a curator may identify an area of low connectivity that might profit from increased curation efforts.

Finally, the properties presented in this paper uncover *semantics* of the data sources and links between sources. Consider the source graph in Fig. 1; there are five alternate paths from Omim to PubMed. Each of these paths yields a different number of distinct objects. Ordering these paths based on the cardinality of objects in PubMed or comparing the overlap of objects among these alternate paths correspond to useful semantics that the researcher can exploit.

In this paper, we develop a model for the source graph paying attention to properties of links and paths and properties of multiple alternate links and paths. We identify properties of the source graph, including the cardinality of objects in a source, the cardinality of objects participating in a link, etc. Given some statistics (for these properties) for a particular object graph, we develop a framework for cardinality estimation in the source graph. This framework allows us to estimate properties of the result graph, i.e., the graph generated as a response to a query that samples the object graph. Finally, we compare the properties of alternate paths, i.e., paths with the same start and end sources but different intermediate sources. This analysis includes comparing the result *cardinality* for each of the alternate paths in some result graph, and the *overlap* of target objects, for pairs of alternate paths. The approach is analogous to database cost models, where statistics

of the database instance are used to predict the result cardinality for a query.

We consider four data sources from NIH/NCBI and the statistics of the corresponding object graph. We sample data from these sources to construct some results graphs, and we validate the accuracy of our framework to estimate the properties of the result graph. Together with related work in [LNRV03] and [LRV03], our research provides a foundation to effectively query and explore data sources.

1.1 Related Work.

There has been prior research on providing access to life science sources [EKL00, ELR01, KRG99, TKM99]. Example systems include DiscoveryLink [HKR+00], Kleisli and its successors [DCB+01], SRS [EA93, EV97] and Tambis [PSB+99]. Recent research in [MSHTH02] has considered multiple alternate paths through sources but they have not addressed the properties of paths. Kleinberg et al. are interested in distinguishing characteristic shapes and connectivity in graphs but not in estimating the number of objects reached etc. as is our interest [Kle99].

Properties of links and paths have been studied in the context of XML document processing in the XSketch project [PG02a, PG02b]. Given an XML document and the corresponding graph, the authors consider label-split graphs and backwards/forwards bi-similar graphs to obtain a synopsis of the original graph. The objective is a compact but accurate synopsis. Assuming statistical independence and uniform distribution, they determine the selectivity for complex path expressions using such synopses. Like us, they use these synopses in an estimation framework. Their approach differs from our approach in that we use statistics such as cardinality and average outdegree from the data graph, rather than detailed synopses.

2 Definitions

This section includes several definitions describing our model of the world and the data within. A logical graph LG with scientific entities as nodes is an abstraction (or schema) of the source graph SG with data sources as nodes. In turn, the object graph OG is an instance of SG. Finally, the result graph RG is a subset of OG and contains the data objects and links specific to a particular query. LG, OG and RG are (somewhat) analogous to the schema, database instance and result of a query.

Definition 1 (Logical Link and Graph) *A scientific entity represents all instances of a class of objects, e.g., gene, sequence, etc. A* logical link *is a directed relationship between two scientific entity classes. The set of scientific entity classes and the logical links between them form the directed* logical graph *LG.*

Queries, such as the one posed in the introduction, choose a single node in the logical graph as a starting point and another single node as an end point. I.e., users start with a certain scientific entity (the starting point) and are interested in its relationship with a certain other entity class (the end point). The purpose of this paper is to explore different paths from starting point to end point by examining the properties of the paths (Sec. 3).

Definition 2 (Source) *A source S is a real-world accessible data source.*

For simplicity of notation, we assume that a source provides data for a single scientific entity class. Thus, in analogy to databases, a source acts as a table. If a real world source provides data for more than one class, we model it as one source for each of its classes. In turn, there can be multiple sources per scientific entity class, leaving users and query planners with certain choices.

Definition 3 (Source link & graph) *A* source link *is a directed edge that connects two sources and corresponds to a logical link. The* source graph *SG is the set of sources and the set of source links.*

For simplicity, we assume that there is only one source per scientific entity. Thus, the logical and source graphs have the same shape. In analogy to databases, source links replace join operations. Instead of keys and foreign keys, sources store links to related objects. Figure 1 shows a source graph with four sources (nodes); each source is annotated with the scientific entity class. Each edge represents a source link. For our example query, the OMIM source is the starting point. To answer the query, OMIM is accessed by directly retrieving objects from that source (using the name of a disease as a keyword). All other sources are accessed by following links.

Definition 4 (Source path) *A source path p is a path from one node (starting point) of the source graph G to another node (end point) of the source graph G. Sources along p are denoted* $S_1^p, \ldots, S_n^p$*, where n is the length of the path.*

Figure 2 lists the five source paths connecting starting point OMIM and end point PUBMED of our example graph.

Definition 5 (Object Link & Graph) *A* object link *is a directed edge between two data objects in two different sources. Each object link implements a source link among the same two sources in the same direction. Given a source graph, the* object graph *OG is a directed graph, in which the set O of all data objects stored by the sources are the nodes, and the set L of object links between these objects are the arcs.*

The object graph represents our model of all the objects and links that we consider. Please note that

- there may be many real links in the sources that are not represented in our model, e.g., a data object could have a link to another data object in the *same* source. Future work will drop the assumption that object links are among different sources.
- a data object can have multiple outgoing and incoming links.
- not all objects have incoming object links.
- not all objects have outgoing object links.
- the graph is not necessarily connected.

Next, we define a result graph as representing the answers (result) of a query against the *OG*. A result graph is a subset of the object graph.

Definition 6 (Result graph; result) *Recall that* $OG = (O, L)$ *is an object graph. Then the result graph* $RG = (O', L')$ *is a graph where* $O' \subseteq O$*,* $L' \subseteq L$*, and* O' *is induced by* L'*.*

Definition 7 (Result path) *A result path RP is a subset of a result graph along a single source path of SG.*

In related work [LNRV03], we have defined a regular expression based query language over the entity classes of *LG*. A regular expression is satisfied by a set of result paths. Each path is a subset of data objects and object links from *OG*, We can use the models developed in this paper to rank paths.

The actual construction of the result graph with a set of real world databases is described in more detail in Sec. 4.1. These graphs were used to test our model.

3 Characterizing the Source Graph

To further our goal of supporting queries on life sciences data sources, we introduce our framework of properties of the source graph such as outdegree, result cardinality, etc. The framework uses statistics from the object graph *OG* such as source cardinality, link cardinality, etc. As we include additional properties and statistics, we further refine the framework and suggest a (refinement to a) formula to estimate the result cardinality of a path.

3.1 Node and Link cardinality

The number of objects stored at a source and their link structure to other sources are among the most basic metadata to obtain, either from the administrators of the sources themselves, or by analyzing source samples. We formally define node and link cardinality for any graph G. Below, we apply these definitions to object graphs and result graphs as defined in the previous sections.

Definition 8 (Node cardinality) *The cardinality $c^G(S)$ of source S is the number of data objects at that source in a graph G. The estimated cardinality $c^G_{est}(S)$ of source S is the estimated number of data objects in that source in a graph G. The set of data objects of S in a graph G is denoted $\{S|_G\}$.*

Definition 9 (Link cardinality) *The* link cardinality *$l^G(S_{i,j})$ denotes the number of links of G from all data objects of source S_i in G pointing to data objects of S_j in G.*

A useful derived property is the average number of outgoing links from data objects:

Definition 10 (Outdegree) *We define the link outdegree $l_{out}(S_{i,j})$ of source S_i as the average number of links of each data object in S_i pointing to an object of source S_j.*

Along a path[1] p, the average outdegree can be calculated as $l_{out}(S^p_{i,i+1}) = l^{OG}(S^p_{i,i+1})/c^{OG}(S_i)$. For brevity, we omit the path index p where the belonging of a source to a path is obvious.

Estimating result cardinality. Let m_1 be the number of starting objects found in source S_1. Following a given path p through sources $S_1, \ldots, S_n$, we construct the result path RP. Assuming independence among object links and no two links pointing to the same object (no overlap), we can calculate the number of distinct objects (i.e., the result cardinality) found of source S_k in RP:

$$c^{RP}_{est}(S_k) = m_1 \cdot \prod_{i=1,\ldots,k-1} l_{out}(S_{i,i+1}), \quad k > 1. \tag{1}$$

The above calculation makes severely simplifying assumptions. The first assumption is link independence along a path. Informally, the probability of a link from an object in source S_{i-1} to an object o in source S_i is independent of the probability of a link from object o in S_i to an object in source S_{i+1}. Future work will examine different dependency cases among links, such as containment, disjointness, etc. The second assumption is link overlap. Informally, two links from source S_{i-1} to S_i may reach one or two *distinct* objects in S_i. We relax this assumption in the following.

To consider overlap of object links, we must determine the likely number of *distinct* objects found in S_k, if randomly choosing m objects from all $c^{OG}(S_k)$ objects in S_k. The probability to find exactly x distinct objects when picking m times from a set of $c^{OG}(S_k)$ objects in a source is (see [Fel68])

$$\frac{\binom{c^{OG}(S_k)}{x} \cdot \binom{m-1}{m-x}}{\binom{m+c^{OG}(S_k)-1}{m}}. \tag{2}$$

For notational simplicity, we define m_k to be the expected number of links from source S_{k-1} to S_k:

$$m_k \coloneqq c^{RP}_{est}(S_{k-1}) \cdot l_{out}(S_{k-1,k}), \quad k > 1.$$

The expected number of distinct objects found in a source is the sum of all possible outcomes x multiplied with their probability from (2):

$$c^{RP}_{est}(S_k) = \begin{cases} m_1, & \text{if } k = 1; \\ \sum_{x=1}^{m_k} x \cdot \frac{\binom{c^{OG}(S_k)}{x} \cdot \binom{m_k-1}{m_k-x}}{\binom{m_k+c^{OG}(S_k)-1}{m_k}}, & \text{if } k > 1. \end{cases} \tag{3}$$

[1] We use path in the usual graph theory sense, i.e., a set of successive directed links through the object graph.

In this formula, we must recursively replace the input value m_k with the number of distinct objects found in the previous source along the path. In our experiments, m_1 is the initial number of records for the starting point source.

Note that we are likely underestimating the overlap due to our independence assumption. In reality, links among semantically related objects are not independent. Most likely they point to a subset of semantically related objects in the other source and link overlap is high. This behavior becomes apparent when we compare the estimated cardinalities with the actual measurements in Sec. 4.

3.2 Image cardinality

The link image of a source link is the set of data objects that are reachable in its implementation.

Definition 11 (Image cardinality) *The link image size $l_{im}(S_{i,j})$ is the absolute number of data objects in S_j that have at least one link pointing to it from parent source S_i in the source graph.*

We are interested in the size of the link image, because this metadata can improve the accuracy of our estimations.

Estimating result cardinality. Including image cardinality into result size estimation modifies the Formula (3) by potentially increasing the overlap. Wherever the cardinality of the source goes into the formula, we replace it with the link image size, since only the objects in the image can participate in RP.

$$c_{est}^{RP}(S_k) = \begin{cases} m_1, & \text{if } k = 1; \\ \sum_{x=1}^{m_k} x \cdot \frac{\binom{l_{im}(S_{k-1,k},S_k)}{x}\cdot\binom{m_k-1}{m_k-x}}{\binom{m_k+l_{im}(S_{k-1,k},S_k)-1}{m_k}} & \text{if } k > 1. \end{cases} \quad (4)$$

with m_k as before for Formula (3).

4 Validating the Framework

We report on an experiment on data sources of the National Center for Biotechnology Information (NCBI) to illustrate that querying well curated sources managed by a single organization may result in different semantics, depending on the specific link, path and intermediate sources that are chosen. Our experiment was limited to the source graph described in Fig. 1; the four data sources hosted at NCBI are NCBI Nucleotide, NCBI Protein, PubMed, and Omim. We sampled data from these sources to construct several results graphs RG. Using our framework, we estimated result cardinality and compared the measured values of RG against our estimates to validate the framework. We also compare the properties of alternate paths.

4.1 Creating Samples and Measurements for *RG*

The methodology to create sample results graphs to validate our framework corresponds to the scenario of retrieving bibliographical references from PubMed that are linked to genes relevant to a given disease or medical condition. We fully explore all links and paths that exist between objects in the four resources, given some start set of objects in Omim.

Consider again the query of Sec. 1: "*Return all citations of* PubMed *that are linked to an* Omim *entry that is related to some disease or condition.*" The study focused on three medical conditions: *cancer*, *aging*, and *diabetes*. For each of these conditions domain experts provided a list of relevant keywords. For example, osteoarthritis, impotence, dietary restriction, maximum work rate, and arthritis characterize the medical condition *aging*. Each set of keywords was used to retrieve relevant genes from Omim via the E-Search utility supported by NCBI. These relevant genes constitute the starting set of objects.

The execution of this query then explored all paths from the starting objects of OMIM, and ended in objects of the end point source PUBMED. We used a wrapper, implemented in Java and Perl, to make successive calls to the E-Link interface provided by NCBI, to follow the links from OMIM records to each of the other 3 sources, as well as to traverse all potential paths (of length 2 and 3) from OMIM to PUBMED.

We created 12 datasets, 4 for each of the conditions. Each dataset starts with a collection of between 140 to 150 OMIM records. The overlap among these OMIM records was very low, and was zero in many cases.

The data was collected in February and March 2003. We note that it took between 20 to 24 hours to download all the data for one set of approximately 150 OMIM records, following the limits on frequency of queries imposed by the NCBI site. We further note that in many cases, there were extremely low levels of time-out errors from E-Link, despite the high volume of requests. These details are of interest since an eventual application of our framework will be in query optimization to reduce query execution time.

Figure 3 shows the results of one experiment for the condition *aging*, starting with 141 OMIM records. It shows the *measured* values for different paths through the result graph. Each edge label shows the link cardinality. For example, there are 1,651 links from the 141 OMIM records to PROTEIN records. Each node label shows the number of distinct objects found by following those links (node cardinality). For example, of the 1,651 links from OMIM to PROTEIN, only 1,590 distinct PROTEIN records are found. We note that outliers were (recursively) eliminated from the *RG*. To identify outliers, we determined the average outdegree for the data sampled in a link, and eliminated those records whose outdegree far exceeded the average. The number of outliers was typically $\ll 1\%$.

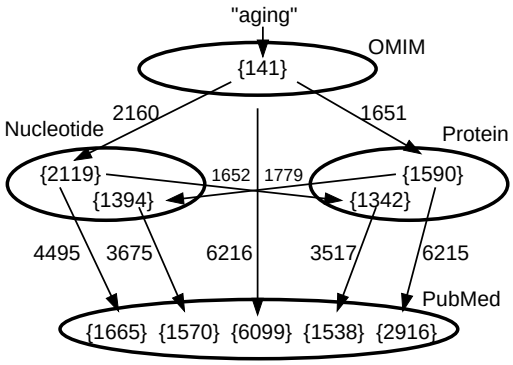

Figure 3: The result graph from experiments on aging

The following table compares the link and node cardinality. The first row shows the link cardinality for the last link of each of the five paths enumerated on Page 2. Note that this number reflects the elimination of outliers[2]. The second row shows the node cardinality (after elimination of duplicates and outliers).

Path	P1	P2	P3	P4	P5
PUBMED entries with duplicates	6216	4495	6215	3517	3675
PUBMED entries without duplicate	6099	1665	2916	1538	1570

This table indicates that the amount of link overlap measured in the five alternate paths differs significantly. Path (P1) has few duplicates (less than 2%) whereas the other paths have more than 50% duplicates (Paths P2 through P5). Such significant variation in the number of duplicates in these paths violates our assumption of independence among the links. It is clear that in those paths with more than 50% duplicates, the probability that an object participates in one link *is not independent* of the probability that this object participates in another link. We return to this assumption when we validate our model.

4.2 Estimations for *RG*

As an input to our formulas, we obtained statistics on the object graph *OG* for February 2003 from NCBI [LL03]. These statistics are summarized in Appendix A. These include node cardinality for each of the sources and link cardinality for any pair of sources. The statistics were used within

[2]An outlier is a record that has a statistically significantly larger number of links compared to the average distribution for the records in that source.

our framework and applied to Formula (4) to estimate the number of distinct objects found at each node.

Figure 4 shows the results of these estimations. The number of OMIM entries (141) in the start node was chosen to exactly correspond to the result graph *RG* of Fig. 3. In Fig. 4 node labels give the estimated number of distinct objects encountered along a path and edge labels give the estimated number of links.

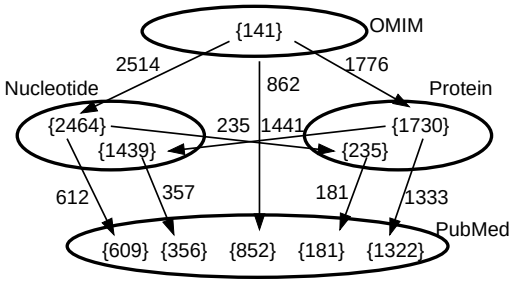

Figure 4: The result graph from calculations

Consider for example the highlighted path OMIM (`Om`) to NUCLEOTIDE (`Nu`) to PUBMED (`Pu`). From the 141 `Om` entries, based on the average outdegree from `Om` to `Nu`, we expect to find 2514 outgoing links to the NUCLEOTIDE data source. Assuming independence of the data objects and links, and knowing the link image of `Nu`, we estimate to find 2464 unique objects of `Nu`. We further estimate the number of outgoing links from `Nu` to `Pu` to be 612 links of which we expect 609 unique objects of `Pu`.

4.3 Comparison

We now compare the estimations in Fig. 4 with the measurements of Fig. 3. To understand the discrepancies, regard Table 1. For each link in the five paths, we report the number of measured links (`LnkMs`), the number of estimated links (`LnkEst`) and the error in estimation as the ratio of estimation and measurement (`LnkEst/LnkMs`). The last three columns are analogous for the number of distinct objects. Note that in our estimation, we use statistics from the object graph.

The error fractions indicate that only for links `Om-Nu`, `Om-Pr` and `Om-Pr-Nu`, the estimates of link cardinality (1.164, 1.076, and 0.810 respectively) and number of distinct objects (1.163, 1.088 and 1.032 respectively) appear to be reasonably close and within some error of approximation. However, if we consider the link `Om` to `Pu`, the error is significant. The estimate are 862 links and 852 distinct objects of `Pu`, whereas the measured values are 6216 and 6099, respectively. This is clearly a gross misestimation.

For links where the error fraction for both links and objects is close to 1.0 (low error), what appears common is that the number of distinct objects is in the same range as the number of links. The independence assumption for links(objects) of our model appears to be upheld here. However, for the rest of the links where the error fraction is close to 0.0 (high error), we observe that the number of distinct objects is significantly lower than the number of links. This indicates that the assumption of an uniform distribution with independence among links(objects) is not supported.

Link	LnkMs	LnkEst	ERROR = LnkEst/LnkMs	ObjMs	ObjEst	ERROR = ObjEst/ObjMs
Om-Pu	6,216	862	0.139	6,099	852	0.140
Om-Nu	2,160	2,514	1.164	2,119	2,464	1.163
Om-Pr	1,651	1,776	1.076	1,590	1,730	1.088
(Om-)Nu-Pu	4,495	612	0.136	1,665	609	0.366
(Om-)Pr-Pu	6,215	1,333	0.214	2,916	1,322	0.453
(Om-)Nu-Pr	1,652	235	0.142	1,342	235	0.175
(Om-)Pr-Nu	1,779	1,441	0.810	1,394	1,439	1.032
(Om-Nu-)Pr-Pu	3517	181	0.051	1,538	180	0.117
(Om-Pr-)Nu-Pu	3,675	357	0.097	1,570	356	0.227

Table 1: Fractional error in estimation for "aging"

5 Training and Testing

Having twelve result graphs RG, one for each set of OMIM starting objects, we enhanced our estimations by training an estimator and testing it. I.e., we used all but one of the result graphs to gain insight into expected path cardinalities given certain input parameters (training). The single remaining result graph served as the test data set, and predictions from the eleven RGs were compared with the actual value of the twelfth RG (testing).

5.1 Model for Training using RGs

For the result graph RG we present some definitions and an expression for our estimation.

Definition 12 (Link cardinality in RG) *$l^{RG}(S_{i,j})$ denotes the number of links from all objects of source S_i in RG pointing to data objects of S_j in RG.*

Definition 13 (Link Participation in RG) *$l^{RG}_{par}(S_{i,i+1})$ is the number of objects in S_i in RG having at least one outgoing link to an object in S_{i+1} in RG.*

Definition 14 (Link Image in RG) *$l^{RG}_{im}(S_{i,i+1})$ is the number of data objects in S_{i+1} in RG that have at least one incoming link from objects in S_i in RG.*

Definition 15 (Outdegree in RG using Participation) *Link outdegree $l^{RG,part}_{out}(S_{i,j})$ of source S_i is the average number of links of each data object in S_i in RG pointing to an object of source S_j in RG.*

Along a path p in RG, average outdegree based on participation is calculated as $l^{RG2}_{out}(S_{i,i+1}) = l^{RG}_{out}(S_{i,i+1})/l^{RG}_{par}(S_{i,i+1})$.

We define *path dependence factor* to capture the statistics from the RGs of an object in S_i having both an inlink from S_{i-1} and an outlink to an object in S_{i+1}.

Definition 16 (Path Dependence Factor(pdf)) *Let $p(i, i+1, i+2)$ be a path of length three. Then $pdf(S_{i,i+1,i+2}) := l^{RG}_{par}(S_{i+1,i+2})/l^{RG}_{im}(S_{i,i+1})$.*

We define *duplication factor* to capture the statistics from the RG of two links from S_i pointing to the same object in S_{i+1}.

Definition 17 (Duplication Factor(df)) *The duplication factor is the ratio of the number of objects in S_{i+1} with an inlink from S_i and the number of outlinks from S_i to S_{i+1}: $df(S_{i,i+1}) := l^{RG}_{im}(S_{i,i+1})/l^{RG}_{out}(S_{i,i+1})$.*

Estimating result cardinality. Let m_1 be the number of starting objects found in source S_1. Following a given path p through sources $S_1, \ldots, S_n$, we construct the result path RP. Recall that assuming independence among object links and no two links pointing to the same object (no overlap), we calculated the number of objects found of source S_k in RP:

$$c^{RP}_{est}(S_k) = m_1 \cdot \prod_{i=1,\ldots,k-1} l_{out}(S_{i,i+1}), \quad k > 1. \tag{5}$$

Using $df(S_{i,i+1})$ and Average Outdegree based on Cardinality $l^{RG1}_{out}(S_{i,i+1})$

$$c^{RP}_{est}(S_k) = m_1 \cdot \prod_{i=1,\ldots,k-1} l^{RG1}_{out}(S_{i,i+1}) \cdot df(S_{i,i+1}), \quad k > 1. \tag{6}$$

Let m_1 be the number of participating objects found in source S_1. Using $pdf(S_{i,i+1,i+2})$ and average outdegree based on participation $l^{RG2}_{out}(S_{i,i+1})$, we can estimate the number of links, the path cardinality, as follows:

$$c^{RP}_{est}(S_k) = m_1 \cdot l^{RG2}_{out}(S_{1,2}) \cdot \prod_{i=2,\ldots,k-1} pdf(S_{i-1,i,i+1}) \cdot l^{RG2}_{out}(S_{i,i+1}), \quad k > 2. \tag{7}$$

We can estimate the number of object, Object Cardinality as follows:

$$c_{est}^{RP}(S_k) = m_1 \cdot l_{out}^{RG2}(S_{1,2}) \cdot df(S_{1,2}) \cdot \prod_{i=2,\ldots,k-1} pdf(S_{i-1,i,i+1}) \cdot l_{out}^{RG2}(S_{i,i+1}) \cdot df(S_{i,i+1}), \quad k > 2. \tag{8}$$

5.2 Validating the Model Using the RGs

For each of the 12 datasets (see Sec. 4.1), outliers were eliminated (recursively) and statistics, such as path dependence factor, duplication factor, average outdegree, etc., were calculated. We then chose two datasets, namely aging1 and diabetes4, to make predictions on by using the average value taken over the remaining 11 datasets. For aging1, the average was calculated over aging2 through diabetes4, and for diabetes4 the average was calculated over aging1 through diabetes3. The result graphs for aging1 and diabetes4 are shown in Fig. 5.

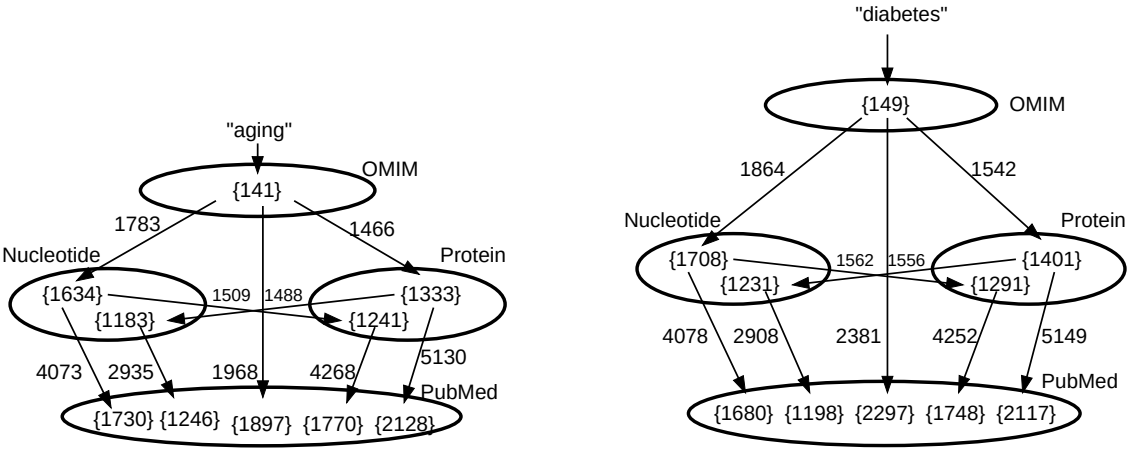

Figure 5: The result graph from predictions on aging1 and on diabietes4

The predictions are compared with the values from experiments and tabulated in Table 2 for aging1 and for diabetes4. The tables are similar to Table 1, where we tabulate errors in estimation assuming independence of links(objects).

For aging1, three of the links (`Om-Nu-Pu`, `Om-Nu-Pr`, and `Om-Nu-Pr-Pu`) have good predictions with the fractions for prediction of links (objects) ranging from 0.91-1.21 (0.92-1.1). Five of them have a moderate prediction for links (objects) with ranges of 0.79-0.88 (0.73-0.84). The only poor prediction is for the link `Om-Pu`, where the error of prediction for the links(objects) is 0.31 (0.31). However, in aging1 the behavior of `Om-Pu` is very different from the other datasets (aging1 for `Om-Pu` has an outdegree of 44.1 while the average is 13.3).

For diabetes4, unlike aging1, most of the values are overpredicted. The link `Om-Pu`, however shows a good prediction for links (objects) with the fraction of prediction for links (objects) being 1.18 (1.16). The only comparatively poor value of prediction is for the link `Om-Pr-Nu-Pu`, where the values are 0.74 (0.53). The prediction values indicate an overall better performance in aging1 than diabetes4. This also emphasizes the difference in natures of aging and diabetes datasets.

6 Further properties

Our framework to model the source graph only considered a limited number of interesting link properties. This will probably impose a limitation on the correctness of estimation using the model. We plan to extend the framework and consider the following properties in future research:

- Link indegree and link outdegree distributions: Our current model assumes uniform distribution of links among the data objects of a source. In reality, there can be significant

"aging1" Link	LnkMs	LnkEst	ERROR = LnkEst/LnkMs	ObjMs	ObjEst	ERROR = ObjEst/ObjMs
Om-Pu	6,216	1,968	0.317	6,099	1,897	0.311
Om-Nu	2,160	1,783	0.825	2,119	1,634	0.771
Om-Pr	1,651	1,466	0.888	1,590	1,333	0.838
(Om-)Nu-Pu	4,495	4073	0.906	1,665	1730	1.039
(Om-)Pr-Pu	6,215	5,130	0.825	2,916	2,128	0.730
(Om-)Nu-Pr	1,652	1509	0.913	1,342	1241	0.925
(Om-)Pr-Nu	1,779	1,488	0.836	1,394	1,183	0.849
(Om-Nu-)Pr-Pu	3517	4268	1.214	1,538	1770	1.151
(Om-Pr-)Nu-Pu	3,675	2,935	0.799	1,570	1,246	0.794
"diabetes4" Link	LnkMs	LnkEst	ERROR = LnkEst/LnkMs	ObjMs	ObjEst	ERROR = ObjEst/ObjMs
Om-Pu	2,016	2,381	1.181	1,966	2,297	1.168
Om-Nu	1,279	1,864	1.457	1,256	1,708	1.360
Om-Pr	947	1,542	1.628	917	1,401	1.528
(Om-)Nu-Pu	2,875	4,078	1.418	1,457	1,680	1.153
(Om-)Pr-Pu	5,789	5,149	0.889	3,002	2,117	0.705
(Om-)Nu-Pr	1,088	1,562	1.436	822	1,291	1.571
(Om-)Pr-Nu	1,095	1,556	1.421	900	1,231	1.368
(Om-Nu-)Pr-Pu	3,911	4,252	1.087	2,148	1,748	0.814
(Om-Pr-)Nu-Pu	3,896	2,908	0.746	2,246	1,198	0.533

Table 2: Fractional errors in prediction for "aging1" and "diabetes4"

variance in the outdegree distribution, and we consider the impact of this variance when validating our model; based on the link distribution, we know that an object that participates in multiple links reaches as many distinct objects as it has links. With this (reasonable) assumption, we can improve our estimations.

- Link dependencies: Seldom are the links along a path independent. The existence of an incoming link into a data object changes the likelihood of an outgoing link to another object. We have already explored this property and plan to further enhance our estimation model to reflect such dependencies.
- Link quality: The information quality of a link's source and target can be used as parameters towards a link quality function. Quality includes parameters like the reputation of a source, whether the link was manually or automatically generated, etc.
- Data source coverage and density: When deciding between alternative paths, a sources size both in number of objects and in number of attributes is of relevance.

Being an additional source of information, the first two properties can serve as input to our prediction model and thus refine it. The latter two properties extend our model in another dimension: Instead of studying the mere cardinalities of results, users are often interested in the quality of the data. In analogy to the cardinality model, a quality model over links can help systems and users compare different paths.

7 Conclusions

The presented research is only a starting point of understanding Web life sciences sources and their relationships with one another. Future work concentrates both on the extension and generalization of the set of properties and on the usage of the presented properties for different scenarios. A list of suggested property extensions was already given in Sec. 6. Additionally, we plan to extend our model by allowing other distributions of links, by including multiple sources for individual scientific entities, and by considering more complex link structures, including cycles and loops.

Despite many experiments on NCBI data sources there is yet much data to explore. Only tight cooperation with domain experts reveals properties and semantics of link structures that are particular to certain sources. For instance, NCBI distinguishes curated links, i.e., links that are generated and checked by humans, and non-curated links, which are generated automatically and are thus of poorer quality. In a next step, we will partition the object graph according to that distinction. For each partition we will determine path properties and compare them. A result of this comparison will give hints on improvement of the automated linking mechanisms.

From these extensions it is a logical next step to use the gained insight to compare links and to compare paths for the applications mentioned in the introduction: Query optimization and data curation. Different semantics, such as path length or result cardinality, can be used to choose the best among several alternative paths through a link structure. Together with results presented in [LNRV03] and [LRV03], this application area promises biologists the ability to efficiently and effectively query interlinked data sources, such as those at NIH/NCBI.

Acknowledgements. This research is partially supported by NSF grants IIS0219909, EIA0130422 and IIS0222847, and the NIH National Institutes of Aging Grant 1 R03 AG21671-01. We thank Barbara Eckman of IBM Life Sciences for insights into source properties, David Lipman and Alex Lash of NCBI for their expertise on NCBI data sources, Damayanti Gupta for data collection, and Marta Janer and Michael Jazwinski for identifying relevant keywords to construct the result graphs.

References

[DCB+01] S. Davidson, J. Cabtree, B. Brunk, J.Schug, V. Tannen, C. Overton, and C. Stoeckert. K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal*, 40(2), 2001.

[EA93] T. Etzold and P. Argos. SRS: An indexing and retrieval tool for flat file data libraries. *Computer Applications of Biosciences*, 9(1), 1993.

[EKL00] B. Eckman, A. Kosky, and L. Laroco. Extending traditional query-based integration approaches for functional characterization of post-genomic data. *BioInformatics*, 17(2), 2000.

[ELR01] B. Eckman, Z. Lacroix, and L. Raschid. Optimized seamless integration of biomolecular data. *Proc. of the IEEE Int. Symp. on Bio-Informatics and Biomedical Engineering*, 2001.

[EV97] T. Etzold and G. Verde. Using views for retrieving data from extremely heterogeneous databanks. *Pacific Symp. on Biocomputing*, pages 134–141, 1997.

[Fel68] W. Feller. *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, New York, NY, 1968.

[HKR+00] L. Haas, P. Kodali, J. Rice, P. Schwarz, and W. Swope. Integrating life sciences data - with a little Garlic. *Proc. of the IEEE Int. Symp. on Bio-Informatics and Biomedical Engineering*, 2000.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Computing Surveys*, 46(5):604–632, 1999.

[KRG99] G. Kemp, C. Robertson, and P. Gray. Efficient access to biological databases using CORBA. *CCP11 Newsletter*, 3.1(7), 1999.

[LL03] A. Lash and D. Lipman. Statistics on NIH/NCBI data sources. *Personal communication*, 2003.

[LNRV03] Z. Lacroix, F. Naumann, L. Raschid, and M.-E. Vidal. Exploring life sciences data sources. In *Workshop on Information Integration on the Web (IIWeb)*, 2003. Joint with IJCAI.

[LRV03] Z. Lacroix, L. Rashid, and M.E. Vidal. Efficient techniques to explore paths in life science data sources. Technical report, University of Maryland, 2003.

[MSHTH02] P. Mork, R. Shaker, A. Halevy, and P. Tarczy-Hornoch. PQL: A declarative query language over dynamic biological data. *Proc. of the AMIA*, 2002.

[PG02a] N. Polyzotis and M. Garofalakis. Statistical synopses for graph-structured XML databases. *Proc. of the ACM SIGMOD Conference*, 2002.

[PG02b] N. Polyzotis and M. Garofalakis. Structure and value synopses for XML data graphs. *Proc. of the Conf. on Very Large Databases (VLDB)*, 2002.

[PSB+99] N.W. Paton, R. Stevens, P.G. Baker, C.A. Goble, S. Bechhofer, and Brass. Query processing in the tambis bioinformatics source integration system. *Proc. of the IEEE Intl. Conf. on Scientific and Statistical Databases (SSDBM)*, 1999.

[TKM99] T. Topaloglou, A. Kosky, and V. Markovitz. Seamless integration of biological applications within a database framework. *Proc. of the Intl. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, 1999.

A Statistics Provided by NCBI

Table 3 presents metadata about four NIH/NCBI data sources [LL03]. For each source we show their cardinality ($c^{OG}(S)$) in the first line. The next lines present cardinalities concerning the source links between two sources: the overall number of links from one source to the other (link cardinality $l^{OG}(S_{i,k})$); the number of data objects having at least one link to the other (link participation $l_{par}(S_{i,k})$); and the number of objects of the other having at least one incoming link from the first source (link image $l_{im}(S_{i,k})$).

Notice that the links in this particular set of sources are symmetric. I.e., for each link in one direction the administrators of the NCBI sources inserted a reverse link. Thus, link participation of one direction and link image of the reverse link are equal. For other sets of sources, in particular if they are maintained by different organizations, this behavior cannot be expected.

	Omim	Nucleotide	Protein	PubMed
Omim	14,759			
$l^{OG}(S_{i,k})$	-	263,129	185,861	90,261
$l_{par}(S_{i,k})$	-	9,863	9,637	13,666
$l_{im}(S_{i,k})$	-	122,826	67,568	73,807
Nucleotide		24,051,882		
$l^{OG}(S_{i,k})$	263,129	-	2,293,022	5,971,098
$l_{par}(S_{i,k})$	122,826	-	1,006,755	5,439,522
$l_{im}(S_{i,k})$	9,863	-	2,040,315	143,599
Protein			2,753,334	
$l^{OG}(S_{i,k})$	185,861	2,293,022	-	2,121,156
$l_{par}(S_{i,k})$	67,568	2,040,315	-	1,560,946
$l_{im}(S_{i,k})$	9,637	1,006,755	-	164,085
PubMed				12,388,558
$l^{OG}(S_{i,k})$	90,261	5,971,098	2,121,156	-
$l_{par}(S_{i,k})$	73,807	143,599	164,085	-
$l_{im}(S_{i,k})$	13,666	5,439,522	1,560,946	-

Table 3: Statistics for four sources