Links and Paths through Life Sciences Data Sources

Zoé Lacroix[1], Hyma Murthy[2], Felix Naumann[3], Louiqa Raschid[2]

[1] Arizona State University
zoe.lacroix@asu.edu
[2] University of Maryland
louiqa,hmurthy@umiacs.umd.edu
[3] Humboldt-Universität zu Berlin
naumann@informatik.hu-berlin.de

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. Analogous to database cost models, we use samples and statistics from the object graph to develop a framework to estimate the result size for a query on the object graph.

1 Querying Interlinked Sources

An abundance of biological data sources contain data about scientific entities, such as genes and sequences. Logical relationships between scientific objects are implemented as *links* between data sources. Scientists are interested in exploring these relationships between scientific objects, e.g., genes and bibliographic 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. Figure 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. In all, there are five paths

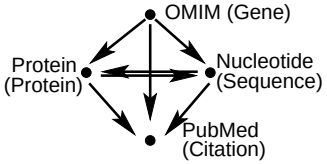

Fig. 1. Source graph for NCBI data sources (and corresponding scientific entities)

(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

Fig. 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. 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. Finally, the properties presented in this paper uncover *semantics* of the data sources and links between sources. Consider the five alternative paths in Fig. 2. 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 while paying attention to properties of links and paths and properties of alternative links and paths. These properties of the source graph allow us to estimate properties of the result graph, i.e., the graph generated as a response to a query against the object graph. 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 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 [10], our research provides a foundation for querying and exploring data sources.

There has been prior research on providing access to life science sources [2, 3, 7, 14]. Example systems include DiscoveryLink [6], Kleisli and its successors [1], SRS [4], and Tambis [12]. Recent research in [11] has considered multiple alternate paths through sources but they have not addressed the properties of paths. In [8] Kleinberg et al. are interested in distinguishing characteristic shapes and connectivity in graphs but not in estimating the number of objects reached, as is our interest.

Properties of links and paths have been studied in the context of XML document processing in the XSketch project [13]. 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 object graph, rather than detailed synopses.

2 Definitions

This section describes our model of the world and the data within. For formal definitions see [9]. In short, 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. 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 an individual source for each of its classes.

The object graph OG represents our model of all the objects and links that we consider. Each object is an instance of a particular class and each link is between two objects of different classes. A data object can have multiple outgoing and incoming links, not all objects have incoming or outgoing object links, and thus the object graph OG is not necessarily connected. 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.

In related work [10], 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. The actual construction of the result graph with a set of real world databases is described in more detail in Sec. 4. 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.

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 define node and link cardinality for any graph G (for formal definitions, we refer to [9]). Then we apply these definitions to object graphs and result graphs as defined in the previous section.

We denote the cardinality of source S as $c^{OG}(S)$, and the estimated cardinality as $c^{OG}_{est}(S)$. We denote the number of links (link cardinality) between sources S_i and S_j as $l^{OG}(S_{i,j})$. A useful derived property is the average number of outgoing links from data objects, calculated along a path[4] p as $l_{out}(S^p_{i,i+1}) = l^{OG}(S^p_{i,i+1})/c^{OG}(S_i)$. The link image of a source is the set of data objects that are reachable in its implementation. Its cardinality is denoted as $l_{im}(S_{i,j})$. We are interested in the size of the link image, because this metadata improves the accuracy of our estimations. 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 and estimate the number of distinct objects reached at the last source of the path, i.e., the result cardinality. To consider overlap of object links, we must determine the likely number of *distinct* objects found in S_i, if randomly choosing m objects from all $c^{OG}(S_i)$ objects in S_i. The probability to find exactly x distinct objects when picking m times from a set of $c^{OG}(S_i)$ objects in a source is (see [5])

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

For notational simplicity, we define m_i to be the expected number of links from source S_{i-1} to S_i, i.e., $m_i := c^{RP}_{est}(S_{i-1}) \cdot l^{RP}_{out}(S_{i-1,i})$ for $i > 1$. The expected number of distinct objects found in a source is the sum of all possible outcomes x multiplied with their probability from (1):

$$c^{RP}_{est}(S_i) = \begin{cases} m_1, & \text{if } i = 1; \\ \sum_{x=1}^{m_i} x \cdot \frac{\binom{l_{im}(S_{i-1},S_i)}{x} \cdot \binom{m_i-1}{m_i-x}}{\binom{m_i+l_{im}(S_{i-1},S_i)-1}{m_i}} & \text{if } i > 1. \end{cases} \tag{2}$$

In this formula, we must recursively replace the input value m_i with the number of distinct objects found in the previous source along the path. This calculation makes the simplifying assumption of link independence along a path. Informally, we assume that the probability of a link from some 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.

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

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. Data was sampled from each of the sources to construct several results graphs *RG*. Validation involved comparing measured values of the *RG*s with our estimates.

Creating Samples and Measurements. The methodology to create sample result graphs corresponds to 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 sources, given the start set of objects in Omim. The study focused on three medical conditions: *cancer*, *aging*, and *diabetes*. A list of relevant keywords for each condition was used to retrieve relevant genes from Omim. These genes constitute the starting set of objects. We created 12 result graphs, 4 for each of the conditions. Each result graph contains a collection of 140 to 150 Omim records and usually many more objects from the other sources.

Figure 3 (left) shows the results of one such experiment for the condition *aging*, starting with 141 Omim records along with the *measured* values for different paths through the result graph. Each edge label shows the link cardinality and each node label shows the number of distinct objects found by following those links (node cardinality).

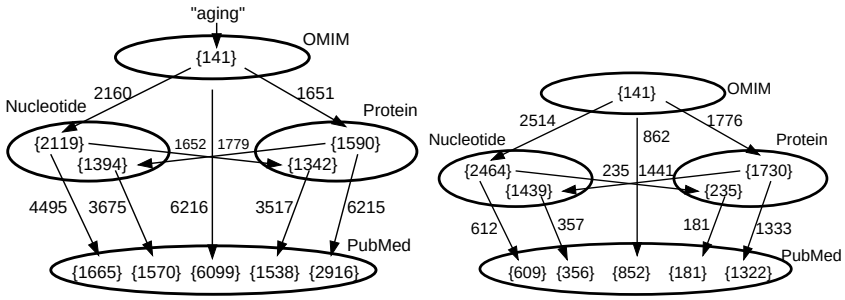

Fig. 3. The result graph from experiments on aging (left) and calculation (right)

Estimations. As an input to our formulas, we obtained statistics on the object graph *OG* for February 2003 from NCBI. These statistics include node cardinality for each of the sources and link cardinality for any pair of sources. The statistics were input to Formula (2) to estimate the number of distinct objects found at each node. Figure 3 (right) shows the results of these calculations. The number of Omim entries (141) in the start node was chosen to exactly correspond to the result graph *RG* of Fig. 3 (left). The node labels on the right give the estimated number of distinct objects encountered along a path and edge labels give the estimated number of links.

Comparison. We now compare the measurements and estimations of the in Fig. 3. To understand the discrepancies, consider Tab. 1. For each link in the five paths, we report the number of measured links (`LinkMeas`), the number of estimated links (`LinkEst`), and the error in estimation as the ratio of estimation and measurement (`LinkEst/LinkMeas`).

Link	LinkMeas	LinkEst	ERROR = LinkEst/LinkMeas	ObjMeas	ObjEst	ERROR = ObjEst/ObjMeas
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"

For those 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 (objects) where the error fraction is close to 0.0 (high error) indicates that the assumption of an uniform distribution with independence among links (objects) is not supported.

5 Training and Testing

Having twelve result graphs RG, one for each set of OMIM starting objects, we enhanced our estimations using a training and testing technique. That is, 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. Through this training, we are able to overcome the independence assumption made in Sec. 3 for result cardinality estimation.

Model for Training. For result graphs RG we present some additional notation and an expression for our estimation. For formal definitions, please see [9]. Link participation $l^{RG}_{par}(S_{i,j})$ is the number of objects in S_i in RG having at least one outgoing link to an object in S_j. Link outdegree in RG using participation (instead of the entire set of object in S_i) is denoted $l^{RG'}_{out}(S_{i,j})$ and describes 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 $l^{RG'}_{out}(S_{i,i+1}) = l^{RG}_{out}(S_{i,i+1})/l^{RG}_{par}(S_{i,i+1})$. To overcome the independence assumption made earlier, we define the *path dependence factor pdf* capturing 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}: $pdf(S_i) \coloneqq l^{RG}_{par}(S_{i,i+1})/l^{RG}_{im}(S_{i-1,i})$. We further define the *duplication factor df* to capture the statistics from the RG of two links from S_i pointing to the same object in S_{i+1}: $df(S_{i,i+1}) \coloneqq l^{RG}_{im}(S_{i,i+1})/l^{RG}_{out}(S_{i,i+1})$.

Estimating result cardinality. Following a given path p through sources $S_1, \ldots, S_n$, we construct the result path RP. Let m_1 be the number of participating objects found in source S_1. Using $pdf(S_i)$ and average outdegree based on participation $l_{out}^{RG2}(S_{i,i+1})$, we can estimate the object cardinality as follows:

$$c_{est}^{RP}(S_k) = m_1 \cdot l_{out}^{RG'}(S_{1,2}) \cdot df(S_{1,2}) \cdot \Pi_{i=2,\ldots,k-1} \left[pdf(S_i) \cdot l_{out}^{RG'}(S_{i,i+1}) \cdot df(S_{i,i+1}) \right], \; k > 2 \qquad (3)$$

Validating the Model. For each of the 12 sampled object graphs (see Sec. 4), statistics, such as path dependence factor, duplication factor, average outdegree, etc., were calculated. We then chose one object graph, namely `aging1`, to make predictions on by using the average value taken over the remaining 11 objects graphs. For `aging1`, the average was calculated over `aging2` through `diabetes4`. The result graph for `aging1` is shown in Fig. 4.

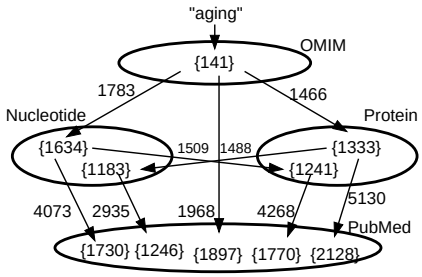

Fig. 4. The result graph from predictions on `aging1`

The predictions are compared with the values from experiments and the results are shown in Tab. 2. The table is similar to Tab. 1, where we tabulate errors in estimation assuming independence of links. Three of the links (`Om-Nu-Pu`, `Om-Nu-Pr`, and `Om-Nu-Pr-Pu`) have good predictions, five of them have a moderate prediction, and the only poor prediction is for the link `Om-Pu`.

"aging1" Link	LinkMeas	LinkEst	ERROR = LinkEst/LinkMeas	ObjMeas	ObjEst	ERROR = ObjEst/ObjMeas
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

Table 2. Fractional errors in prediction for `aging1`

6 Conclusions

The presented research is only a starting point of understanding Web-based 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. Additionally, we plan to extend our model by allowing other distributions of links (stored as histograms), by including multiple sources for individual scientific entities, and by

considering more complex link structures, including cycles and loops. Together with results presented in [10], this application area promises biologists the ability to efficiently and effectively query interlinked data sources, such as those at NCBI.

Acknowledgements. This research is partially supported by NSF grants IIS-0219909, EIA0130422 and IIS0222847, and the NIH National Institutes of Aging Grant 1 R03 AG21671-01. We thank Barbara Eckman of IBM Life Sciences, David Lipman and Alex Lash of NCBI, Damayanti Gupta, Marta Janer, and Michael Jazwinski.

References

1. 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.
2. B. Eckman, A. Kosky, and L. Laroco. Extending traditional query-based integration approaches for functional characterization of post-genomic data. *BioInformatics*, 17(2), 2000.
3. 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.
4. T. Etzold and P. Argos. SRS: An indexing and retrieval tool for flat file data libraries. *Computer Applications of Biosciences*, 9(1), 1993.
5. W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, New York, NY, 1968.
6. 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.
7. G. Kemp, C. Robertson, and P. Gray. Efficient access to biological databases using CORBA. *CCP11 Newsletter*, 3.1(7), 1999.
8. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Computing Surveys*, 46(5):604–632, 1999.
9. Zoé Lacroix, Hyma Murthy, Felix Naumann, and Louiqa Raschid. Links and paths through life sciences data sources. Technical Report, Humboldt-Universität zu Berlin, Institut für Informatik, 2004.
10. Zoé Lacroix, Louiqa Raschid, and Maria-Esther Vidal. Efficient techniques to explore and rank paths in life science data sources. In *Proc. of the Int. Workshop on Data Integration for the Life Sciences (DILS)*, Leipzig, Germany, 2004.
11. P. Mork, R. Shaker, A. Halevy, and P. Tarczy-Hornoch. PQL: A declarative query language over dynamic biological data. *Proc. of the AMIA*, 2002.
12. 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.
13. N. Polyzotis and M. Garofalakis. Structure and value synopses for XML data graphs. *Proc. of the Conf. on Very Large Databases (VLDB)*, 2002.
14. 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.