An Interactive Data Repository with Visual Analytics

Ryan A. Rossi
Palo Alto Research Center
rrossi@parc.com

Nesreen K. Ahmed
Intel Labs, Intel Corp.
nesreen.k.ahmed@intel.com

ABSTRACT

Scientific data repositories have historically made data widely accessible to the scientific community, and have led to better research through comparisons, reproducibility, as well as further discoveries and insights. Despite the growing importance and utilization of data repositories in many scientific disciplines, the design of existing data repositories has not changed for decades. In this paper, we revisit the current design and envision interactive data repositories, which not only make data accessible, but also provide techniques for interactive data exploration, mining, and visualization in an easy, intuitive, and free-flowing manner.

Categories and Subject Descriptors

G.2.2 [**Graph theory**]: Graph algorithms; H.2.8 [**Database Applications**]: Data Mining; H.3.3 [**Information Storage and Retrieval**]: Relevance feedback; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

Keywords

Data repository, data archive, digital library, cyberinfrastructure, interactive data repository, visual analytics, interactive visualization, interactive graph mining, graph visualization, network science, sensemaking, network repository

1. INTRODUCTION

Scientific progress often relies on standard data sets for which claims, hypotheses, and algorithms can be compared and evaluated. In recent years, scientific data repositories have made data widely accessible to the broader scientific community, and have led to better research practices through comparisons, reproducibility, as well as further discoveries and innovations. Such data repositories are proving to be increasingly valuable to many scientific disciplines (e.g., computer science, bioinformatics, etc.) [6; 3], while other disciplines have only recently considered data sharing (e.g., ecology, evolutionary biology, and psychology) [15; 9]. Furthermore, the recent hype of big data has fueled the importance of sharing data for the greater good (e.g., healthcare, climate change). Hence, sharing data and making it accessible is quickly becoming a standard, and in many disciplines is now a requirement for funding [7; 8]. All of these reasons have led to the growing number of data repositories and their widespread use across a variety of disciplines.

Despite the growing importance and utilization of data repositories in many scientific disciplines, the design of existing data repositories has not changed for decades. Most existing data repositories are designed for data sharing and management rather than for scientific inquiry, which impedes the possibility to easily explore the data and ask novel questions beyond the questions that sparked data collection. This is due to the current design of data repositories, which lacks interactive visual analytics [13; 5; 2], mining, and statistical tools that make it easier to understand, explore, and find new and important patterns in the data.

In this paper, we revisit the traditional data repository concept that has been widely used for decades, and instead, we envision *interactive data repositories* (iDR) [10; 11], an alternative approach for the design of future data repositories, which not only makes data accessible, but also provides techniques and tools to find, understand, and explore data in an easy, intuitive, and free-flowing manner. Interactive data repositories combine interactive visualizations with analytic techniques to reveal important patterns and insights for sense-making, reasoning, and decision-making. In addition, they facilitate research, education, training, and scientific discovery. These repositories allow the user to explore a

Figure 1: Network Repository (NR) is a data repository with interactive data exploration and visualization. NR is accessible online at http://networkrepository.com. NR goes beyond data sharing and accessibility by providing state-of-the-art *visual analytic techniques* for real-time interactive data exploration/mining *and* visualization.

single data set, or compare and contrast multiple data sets. In particular, interactive data repositories integrate visual analytic tools, which give the user full control to explore and understand the data in real-time. Data can be explored and visualized through user-defined and free-flowing transformations, queries, filtering, among other possibilities. We posit that the proposed interactive data repository concept will replace existing data repositories that have been used for decades.

We argue that interactive data repositories will significantly speedup scientific progress and discovery by making data accessible, but more importantly, by making data more discoverable, interpretable, and reusable. This would provide the broader scientific community with tools to quickly validate research findings, helping the peer-review process, and understand the caveats of published approaches based on the data and its characteristics. These tools would make it easier and more intuitive to explore the data in real-time, without the overhead of downloading the data, formatting, writing code/loading it, among others.

The remainder of this paper is organized as follows: Section 2 proposes an interactive data repository for graphs and provides a prototype, whereas Section 3 investigates independent and identically distributed (IID) data. Finally, Section 4 concludes.

2. INTERACTIVE GRAPH REPOSITORY

This section discusses the design of an interactive data repository for graphs (a.k.a relational data, networks), where the nodes represent entities (e.g., objects, people) and the links represent the dependencies among them. Graphs arise as a natural data representation, and allow us to study phenomena in a variety of domains, including social, behavioral, biological, transportation, communication, and financial domains. Studying these real-world graphs is crucial for solving numerous problems that lead to high-impact applications. For example, identifying the behavior and interests of users in online social networks (e.g., viral marketing, online advertising), monitoring and detecting virus outbreaks in human contact networks, predicting protein functions in biological networks, and detecting anomalous behavior in computer networks.

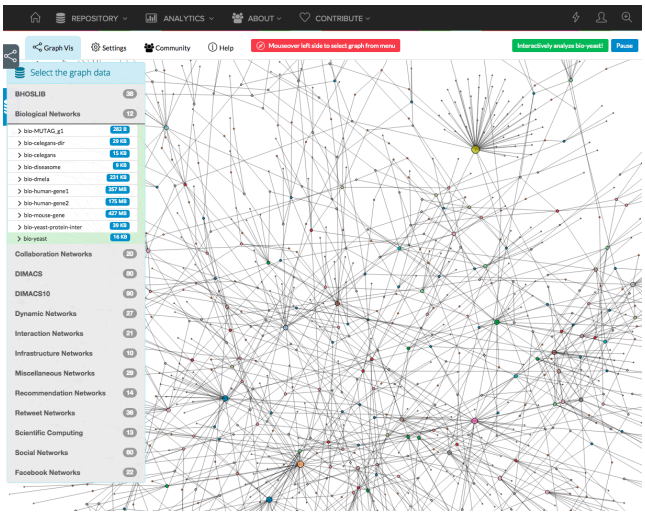

Figure 2: Interactive exploration and visualization of the properties and topology of graphs using Network Repository (NR) to discover valuable insights.

For demonstration, we discuss Network Repository (NR[1]) — the first graph data repository with a web-based interactive platform for real-time graph analytics (Figure 1). NR has hundreds of graphs for users to download and share. However, the key factor that differentiates NR from other repositories [12; 14] is the interactive graph analytics and visualization platform.

Network repository aims to improve and facilitate the scientific study of graphs by making it easy to interactively explore, visualize, and compare a large number of graphs across many different dimensions and facets. NR currently has 500+ graphs from 19 general collections (social, information, and biological networks, among others) that span a wide range of types (e.g., bipartite, temporal) and domains (e.g., social science, physics, biology). In addition to exploring the data in the repository, we also make it easy for users to upload and quickly explore and visualize their own data using the platform.

Next, we discuss some of the key features that that differentiate NR from other repositories.

2.1 Interactive Graph Topology Visualization

The interactive platform gives users the unique ability to interactively explore and visualize the topology of graphs in seconds. Figure 2 demonstrates this feature, where users have the flexibility to visualize any graph in the repository by simply selecting it from the left menu. The left menu displays a variety of graph collections which users can then click to display all graphs in a given collection. Once a graph is selected, we can then get a global view of the structural patterns by zooming-out completely. Similarly, users can drill-down on the regions of the graph that are of interest. For instance, suppose a user is interested in large cliques, then after spotting such regions from the global view, they can zoom into these regions to obtain additional information on the members of the clique and their connections and graph characteristics.

2.2 Multi-scale Interactive Graph Analytics

In order to provide the most flexibility for exploring data, NR provides a multi-scale graph analytics engine. This allows for each graph property to be easily analyzed at various levels of granularity and aggregation, which leads to a large space of possibilities for exploring and querying the data. Such an approach has many other advantages beyond providing users with a large space of possibilities for exploring and querying the data. In particular, NR provides an intuitive and meaningful approach that facilitates exploring and understanding graphs and their structure, both at the global macro-level as well as the local micro-level. For instance, at the global macro-level, NR maintains a number of global graph statistics and properties (e.g., total number of triangles, average clustering coefficient, max k-core number, etc). Alternatively, NR uses node-level (link-level) graph properties to explore graphs at the local micro-level. In addition, the multi-scale analytics engine leverages visual analytics tools that facilitate graph exploration. For

[1] http://networkrepository.com

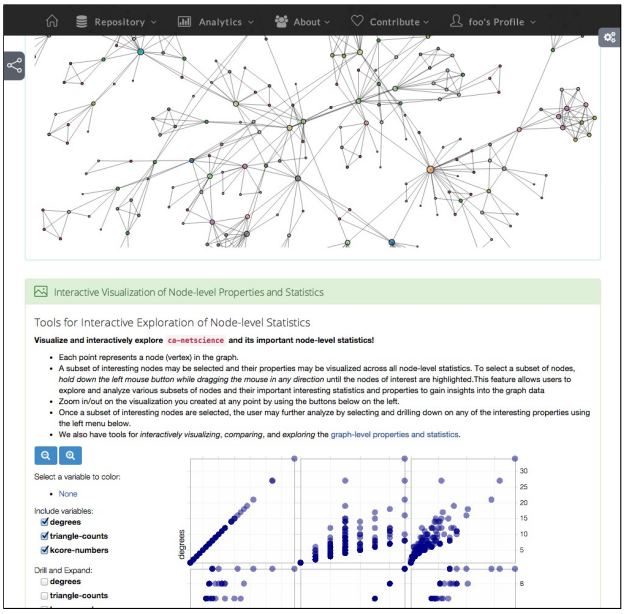

Figure 3: A snapshot of the online page of the network science coauthorship graph (ca-netscience), showing the interactive graph visualization and scatter plot. Note that each graph is automatically processed and assigned a unique URL for reference. This URL makes it easy for others to download the exact data, but also contains documentation and metadata, as well as numerous interactive visualization tools, graph statistics, as well as node-level statistics and distributions.

example, an interactive scatter plot matrix to analyze the correlation between pairs of node/link statistics (see an example in Figure 3), which supports brushing to allow users to highlight interesting nodes (and links) across the various measures. Furthermore, semantic zooming can be used to drill-down in order to understand the differences between individual nodes and links.

Further, NR leverages node and link summarization techniques (e.g., binning/histograms, statistical distributions) to obtain fast, meaningful and useful data representations. For instance, NR provides interactive plots of the cumulative distribution function (CDF) and the complementary CDF for important graph properties (e.g., degree distribution). These are known to be important for networks, capturing interesting structural properties such as heavy-tailed distributions (see an example in Figure 4).

2.3 Interactive Graph Search & Comparisons

Graphs are easily compared across a wide range of important and fundamental graph statistics and properties (e.g., max k-core number, total number of triangles, degree, max clique size, motif counts, etc.). Figure 5 demonstrates how graphs can be interactively compared using an interactive scatter plot matrix and gives intuition for the types of queries and questions that can be explored. Clearly, as we show in Figure 5, there is a collection of data points where each point represents a graph, and users can use brushing to filter graphs via any user-selected constraint(s) and then highlights all such graphs (or nodes/edges) that satisfy it across all other interactive plots. In essence, NR supports interactive techniques such as brushing, linking, highlighting, as well as semantic zooming, to give the user full control to explore, query, and compare large collections of graphs across many dimensions. Finally, NR provides search tools to search for graphs by keywords and types.

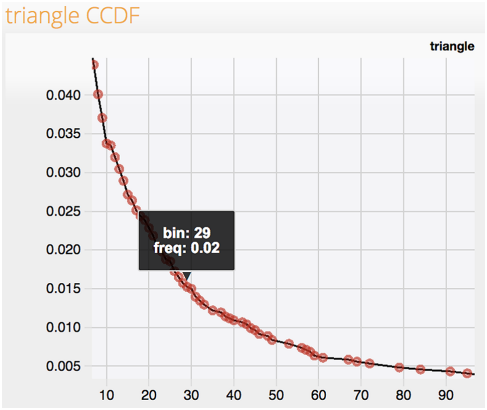

Figure 4: Interactive plot of the triangle count complementary cumulative distribution function (CCDF)

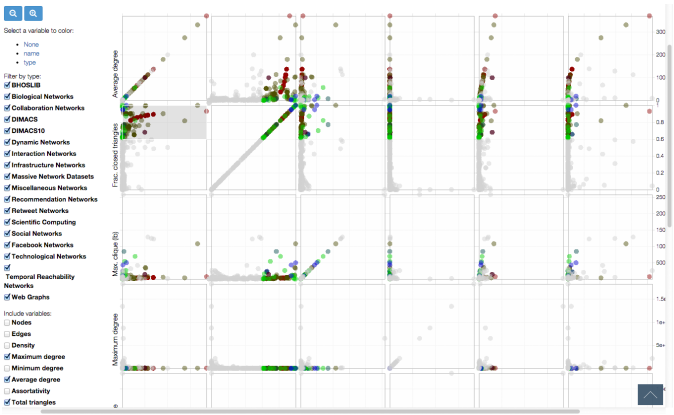

Figure 5: **Interactive scatter plot matrix for large-scale graph comparisons**. Interactively comparing graphs across a wide range of fundamental graph properties. Each data point represents a graph and each unique color represents the graph collection (e.g., social networks). In this example, we filter all graphs that have a global clustering coefficient (κ) greater than 0.6. Thus, all graph datasets that satisfy this query are highlighted in all other interactive plots. Further queries and research questions may be explored using this set of graphs that satisfy $\kappa \geq 0.6$.

Attribute	Range	Mean	Mode	Median	Variance	Skewness	Kurtosis	Median Dev.	Mean Dev.	Coeff. Var.	Missing values
sepal length	3.6	5.84	5	5.8	0.68	0.31	-0.57	0.68	0.69	0.14	0
sepal width	2.4	3.06	3	3	0.19	0.32	0.18	0.33	0.34	0.14	0
petal length	5.9	3.76	-	4.35	3.1	-0.27	-1.4	1.49	1.56	0.47	0
petal width	2.4	1.2	0.2	1.3	0.58	-0.1	-1.34	0.64	0.66	0.63	0

Figure 6: Univariate statistics are updated on-the-fly after any data filtering or querying/selection.

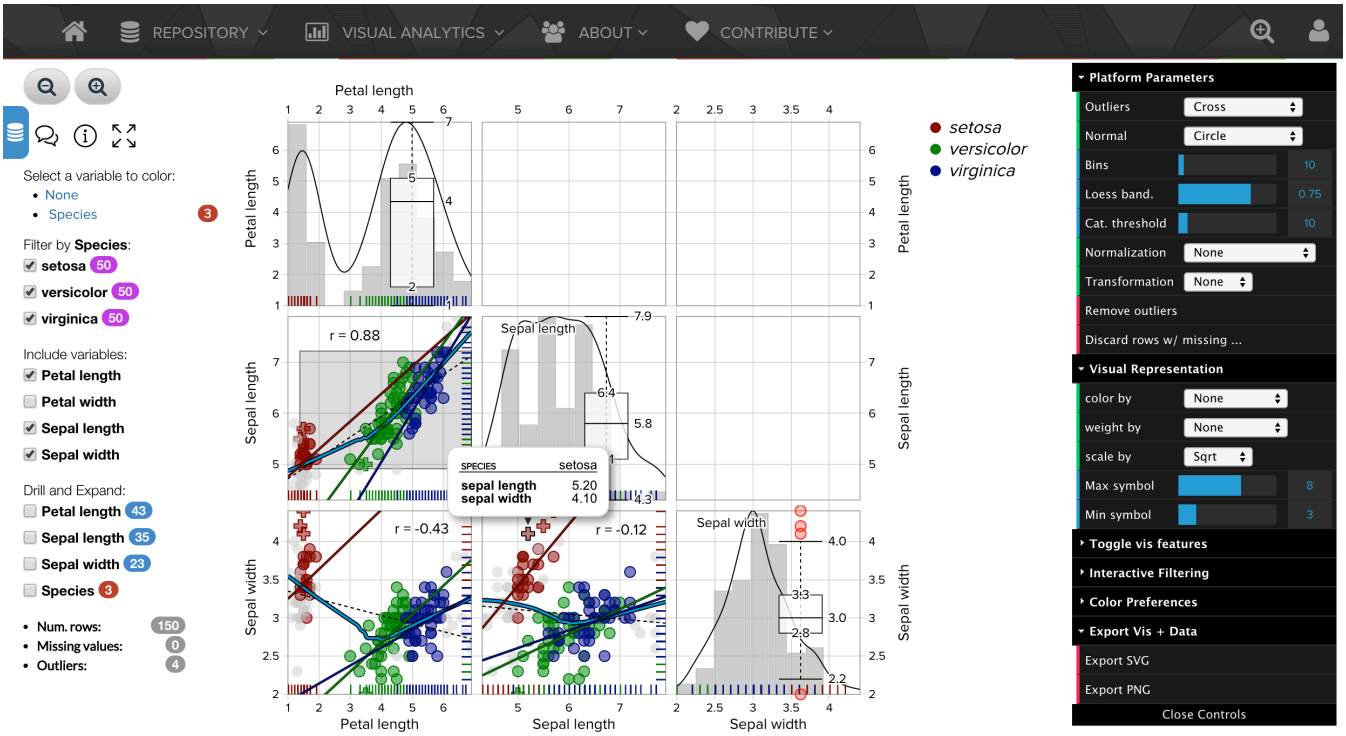

Figure 7: Interactive scatter plot matrix. The screenshot above is of the iris data where data points are colored by species. The lower triangular shows the relationship between pairs of variables, whereas the diagonal provides univariate analysis (e.g., histograms, whisker plots, and outliers).

2.4 Scalability & Big Data Considerations

Big graph data may also be interactively explored and visualized using NR. We don't just provide users with summary or graph-level statistics, but allow a much deeper exploration of the data while sending a significantly smaller amount of data. For instance, users can interactively explore a range of distributions from a wide variety of important graph properties and statistics. Whenever necessary, we utilize state-of-the-art graph sampling methods to ensure fast and efficient loading and processing of the data while being as accurate as possible [1]. These techniques are extremely effective for sampling node features and visualizing the structure and connectivity of the graphs.

Furthermore, at the heart of the interactive platform lies a high-performance parallel graph analytics engine, which is written in C/C++ and designed to be fast and scalable for extremely large graphs. We note that it outperforms other libraries such as GraphLab and igraph (e.g., on triangle and motif counting).

3. STATISTICAL VISUAL ANALYTICS

This section discusses the design of interactive data repositories for IID data. Since visual analytic techniques for iDR largely depend on the scientific discipline, we discuss general guidelines and provide examples from a recent visual analytic platform for such data.

Interactive univariate analysis offers a quick assessment of a variable (e.g., see Figure 6 for point statistics). Further, Figure 7 provides interactive box-and-whisker plots, histograms, outlier detection/visualization, etc. To quantify the relationship between two variables, one may use bivariate point statistics (e.g., correlation coeff. denoted by r in Figure 7). A variety of visual bivariate analytic techniques are shown in the lower-triangular region of Figure 7. In particular, interactive scatter plots, loess curves (non-parametric, non-linear) [4], and regression lines. Categorical variables may be used to color the data points in each scatter plot as well as the loess curves and regression lines.

It is also important to provide interactive multidimensional analytic techniques to understand the relationships between variables simultaneously (e.g., the interactive scatter plot matrix in Figure 7 with brushing and linking).

Interaction techniques such as brushing, linking, zooming, panning, filtering are used heavily in iDR. All data normalization and transformations in Figure 7 are interactive, rapid, incremental, and reversible.

4. CONCLUSION

This paper revisits existing scientific data repositories, and instead, proposes the concept of an *interactive data repository* that aims to facilitate scientific progress by incorporating interactive visual analytic techniques for the exploration, mining, and understanding of data in real-time. The paper also discusses a prototype of interactive data repositories for both graph data and independent and identically distributed (IID) data.

Acknowledgments

We thank all donors who contributed data to the repository and all others who have supported and continue to support this effort.

5. REFERENCES

[1] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: From static to streaming graphs. *Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):7:1–7:56, June 2014.

[2] N. K. Ahmed and R. A. Rossi. Interactive visual graph analytics on the web. In *ICWSM*, pages 566–569, 2015.

[3] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, et al. Arrayexpress–a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, 31(1):68–71, 2003.

[4] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[5] D. Ebert, K. Gaither, Y. Jang, and S. Lasher-Trapp. Cross-scale, multi-scale, and multi-source data visualization and analysis issues and opportunities. In *Scientific Visualization*, pages 353–360. 2014.

[6] P. Murphy and D. W. Aha. UCI repository of machine learning databases–a machine-readable repository. 1995.

[7] NSB. Digital research data sharing and management. Technical Report NSB-11-79, National Science Foundation, 2011.

[8] N. A. of Sciences. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. The National Academies Press, 2009.

[9] H. Pashler and E.-J. Wagenmakers. Editor's introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530, 2012.

[10] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] R. A. Rossi and N. K. Ahmed. Networkrepository: An interactive data repository with multi-scale visual analytics. In *arXiv:1410.3560v2*, 2015.

[12] SNAP. Stanford network dataset collection. http://snap.stanford.edu/data/index.html.

[13] J. J. Thomas, K. Cook, et al. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[14] UCI ML Repository. UCI machine learning repository. http://archive.ics.uci.edu/ml.

[15] M. C. Whitlock, M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. Data archiving. *The American Naturalist*, 175(2):145–146, 2010.