NetworkRepository: A Graph Data Repository with Visual Interactive Analytics

Ryan A. Rossi
Dept. of Computer Science
Purdue University
rrossi@purdue.edu

Nesreen K. Ahmed
Dept. of Computer Science
Purdue University
nkahmed@purdue.edu

ABSTRACT

NetworkRepository.com (NR) is the first *interactive data repository* with a web-based platform for visual interactive analytics. Unlike other data repositories (e.g., UCI ML Data Repository, and SNAP), the network data repository (networkrepository.com) allows users to not only download, but to interactively analyze and visualize such data using our web-based interactive graph analytics platform. Users can in real-time analyze, visualize, compare, and explore data along many different dimensions. The aim of NR is to make it easy to discover key insights into the data extremely fast with little effort while also providing a medium for users to share data, visualizations, and insights. Other key factors that differentiate NR from the current data repositories is the number of graph datasets, their size, and variety. While other data repositories are static, they also lack a means for users to collaboratively discuss a particular dataset, corrections, or challenges with using the data for certain applications. In contrast, we have incorporated many social and collaborative aspects into NR in hopes of further facilitating scientific research (e.g., users can discuss each graph, post observations, visualizations, etc.).

Categories and Subject Descriptors

G.2.2 [**Graph theory**]: Graph algorithms; H.2.8 [**Database Applications**]: Data Mining

Keywords

interactive data repository, data archive, interactive graph mining, graph visualization, network analysis, interactive graph generation

1. INTRODUCTION

This paper presents NetworkRepository.com (NR) — the first data repository with a web-based interactive platform for real-time graph analytics. NR has hundreds of graphs and network datasets for users to download (and share). However, the key factor that differentiates NR from other repositories [5, 6] is our interactive graph analytics and visualization platform. NR allows users to interactively, in real-time, explore and visualize the data.

Scientific progress depends on standard datasets for which claims, hypotheses, and algorithms can be compared and evaluated. NetworkRepository.com aims to improve and facilitate the scientific study of networks and other data by making it easy to interactively explore, visualize, and compare a large number of datasets. NR is the first *interactive graph data repository* that provides researchers with the ability to interactively explore and visualize data in seconds using our fast and easy-to-use interactive analytics platform (e.g., Figure 1 and 2). The repository has a comprehensive and representative set of the most popular and frequently used datasets in academia and industry. More specifically, NR currently has 500+ graphs from 19 general collections (social, information, and biological networks, among others) that span a wide range of types (bipartite, time-series, etc.) and domains (social sciences, physics, bioinformatics).

Unlike other data repositories (e.g., UCI ML Data Repository [6], SNAP [5]), NR allows users to not only download, but to interactively analyze and visualize the data in real-time on the web (e.g., see Figure 2). NR goes beyond traditional static repositories by giving users the ability to interactively explore and compare data along many different dimensions and facets. The goal of NR is to make it easy for users to discover key insights into the data extremely fast with little effort while also providing a medium for researchers to share data, visualizations, and insights. 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.

Static plots found in papers and other repositories are severely limiting as they only provide a single view of the data. By contrast, the interactive platform gives rise to an infinite number of possible views (e.g., scaling, zooming, filtering, and other data transformations). Thus, NR gives researchers the freedom and flexibility to interactively plot and visualize the data according to the properties and characteristics of interest to them. Researchers can begin analyzing and investigating the data independently, asking their own questions, and/or verifying recently published findings/claims. For instance, users can zoom-in on interesting data points (e.g., nodes and/or graphs) as well as scale the data (linear, log, exp, etc.) for specific applications and/or questions being explored.

The platform also allows researchers to easily explore, analyze, and compare graph data in an interactive fashion by selecting (or filtering) data points (representing graphs, nodes, and/or edges) across a variety of important and fundamental graph statistics and properties (see Figure 4). Intuitively, this filtering and selection tool highlights all such nodes that have certain properties of interest such as the nodes that have a triangle count between a certain user-defined range. Thus, NR's interactive platform gives rise to an infinite amount of ways to visualize and compare such data in a real-time web-based platform that is easy and in-

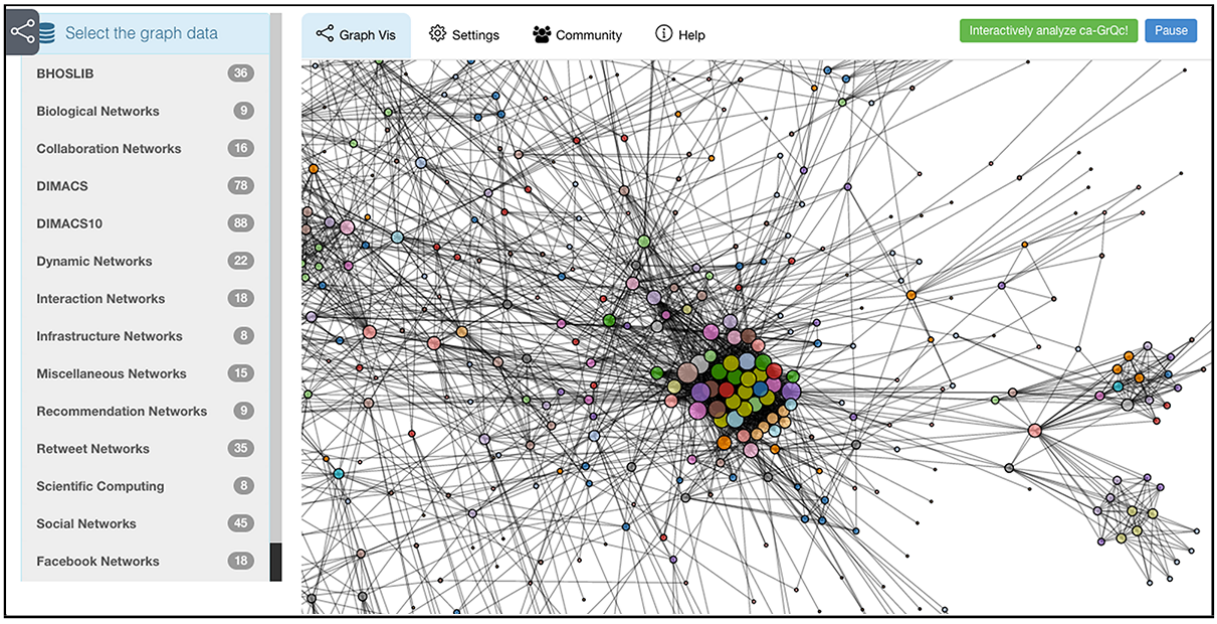


Figure 1: Visualize graph structure/connectivity and discover valuable insights using our interactive graph visualization platform. Compare with hundreds of other networks across many different collections and types. A zoomed-in snapshot of the collaboration network ca-GrQc. Note the links represent scientific collaborations between authors who submitted papers to "general relativity" and the "quantum cosmology" category.

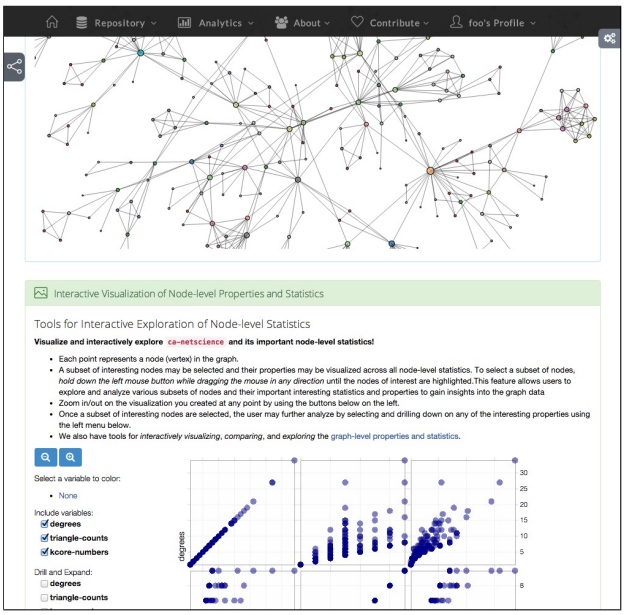


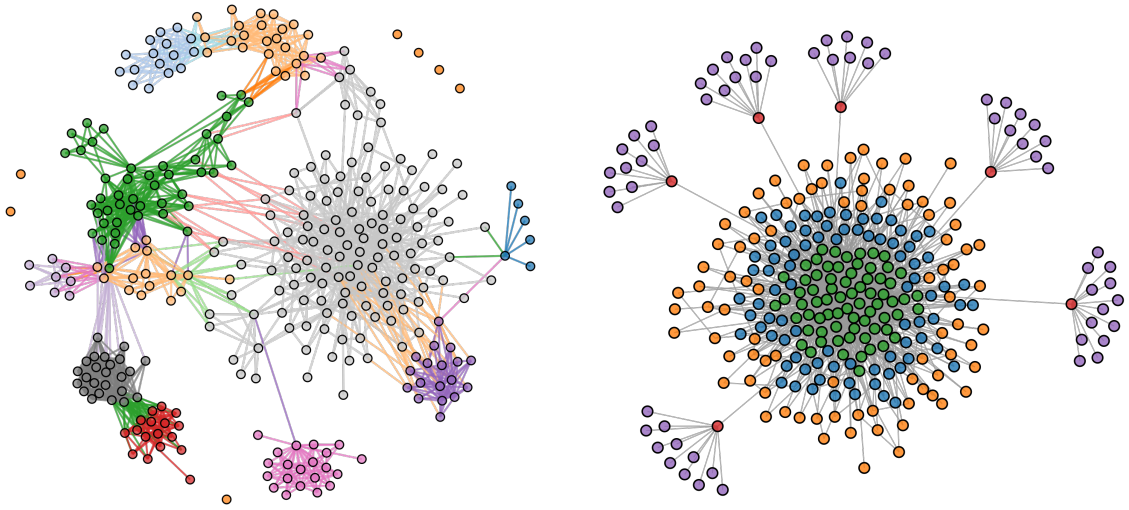
Figure 2: A snapshot of a graph's page showing the interactive graph structure visualization and node-level statistics for ca-netscience. Note that each graph is automatically processed and assigned a unique URL for reference. This URL makes it easy for others to download the exact data, but also contains documentation and metadata, as well as numerous interactive visualizations of the graph structure and connectivity, graph-level/point statistics, as well as node-level statistics and distributions.

tuitive to use.

The interactive data analytics platform is flexible and has many potential applications and use cases. For instance, it has shown to be useful for tasks such as spotting anomalous nodes and subgraphs through interactive comparisons across a wide range of fundamental graph properties and features. Furthermore, we also provide many other interactive analysis tools, e.g., *interactive graph clustering tasks* such as role discovery and community detection (see Figure 3).

Despite the increasing interest in graph data and algorithms, there still remains a lack of standard benchmark datasets for many problems and research areas. Unfortunately, most research uses proprietary data and/or some preprocessed versions of existing network datasets. Thus, it is often impossible to find the original data used in published experiments, and at best it is difficult and time consuming. However, data with ambiguities (e.g., data with the same name) can easily be resolved and identified using NR through via the interactive platform. For the purpose of reproducible research, we encourage users to upload data (including a reference to the published paper), even if the data has been preprocessed for a particular problem/domain. Thus, users can leverage NR to quickly find and understand the data of interest to them, even if the name and other properties are ambiguous (as compared to other repositories where the data must first be downloaded, processed, etc.). In addition, NR is a community-oriented repository that allows users to discuss, share observations, recent findings/papers, and any other insights. This may help provide standardized benchmark graph data, while facilitating comparisons of various algorithms and models. We summarize a few of the contributions and features below.

- An interactive data repository where researchers can quickly compare, explore, and analyze over 500+ graphs interactively in real-time via NR's web-based platform.



(a) Communities (b) Roles

Figure 3: Visual and Interactive Graph Clustering. (a) NR also gives users the ability to visualize and interactively discover communities in any arbitrary graph via a fast community detection designed for interactive exploration. In this visualization, nodes and edges are colored by the community labels. (b) Similarly, *graph roles* may also be discovered and interactively explored. Nodes are colored by their graph role and are shown to be intuitive and easily explainable. For instance, nodes representing star edges are assigned a distinct role while star-center nodes are assigned a different yet consistent role. Chung-Lu graphs also appear to have only a few distinct roles that are easily characterized by relative degree.

- Interactive visualization and exploration of the graph structure (e.g., nodes and edges).
- Global network statistics and parameters (e.g., triangle counts, average. clustering coefficient, maximum k-core number, etc) can be interactively analyzed, visualized, and compared among graphs.
- Local node-level network statistics and features (e.g., k-core number of each node, number of triangles incident to each node, etc).
- Interactive visualizations and plots of key statistical distributions of each network (e.g., degree distribution).
- Community-oriented data repository where users can create profiles (see Figure 5), donate datasets, share visualizations, and insights, as well as save their synthetically generated networks and visualizations created using NR.
- Upload your own data and use our platform to analyze and visualize it.
- Model-based synthetic graph generation and visualization, using standard models such as Erdos-Renyi, Chung-Lu, and preferential attachment.
- Pattern-based synthetic graph generation and visualization, using subgraph patterns such as nodes, edges, cliques, stars, cycles, and chains.
- Hybrid synthetic graph generation and visualization that allows users to generate graphs using a standard model (such as Erdos-Renyi) in addition to adding certain patterns to the generated graph (e.g., cliques, and stars).

2. INTERACTIVE GRAPH REPOSITORY

While NetworkRepository.com is the first interactive graph and network data repository, it also encompasses a wide variety of network data useful for a range of network analysis and application areas. Overall, NR provides interactive filtering of data directly and instantly, incorporates multiple comparative plots of a single statistic as well as across a wide range of statistics, and also provides multiple complementary perspectives that can be viewed simultaneously. Finally, whenever appropriate we allow the user to select subsets of the data which are then highlighted instantly and automatically across the multiple views and statistics. This section discusses in greater detail a few of the fundamental advantages of NR over the current existing data repositories.

2.1 Facilitating Evaluation and Research

NR serves as a testbed to enable researchers interested in network analysis, relational learning, graph mining, knowledge discovery (among others) to test and compare their algorithms on a wide variety of network benchmark datasets. We conjecture NR could lead to improvements in the evaluation of graph/network algorithms (e.g., machine learning, descriptive modeling methods) by providing researchers a large number of data for which their methods can be systematically compared and evaluated. Thus, researchers would better understand what types of data are suitable for certain algorithms/techniques. Ultimately, this would lead to more statistically significant research and findings by providing a large number of graphs from a variety of domains for evaluating novel graph algorithms.

2.2 Multi-level Graph Statistics and Features

In order to provide the most flexibility for exploring data, we take a multi-level approach that allows for each statistic to be analyzed at various levels of granularity and aggregation. This approach also has many other advantages beyond providing users with a large space of possibilities for exploring and querying the data. For example, graph statistics (and features) computed over big data may be compressed via distributions and then used for exploration purposes. In particular, we currently use local node-level (and edge-level)

network stats and features (e.g., k-core number of each node, number of vertex triangles formed at that node, and numerous others). Further, we also provide interactive plots of the important network distributions as well as CDF and CCDF plots for each graph property/feature as well as global network statistics and parameters (e.g., total triangles, avg. clustering coefficient, max k-core number, etc).

2.3 Visualizing Graph Structure

The platform also gives users the unique ability to interactively explore and visualize the structure and connectivity of graph data in seconds. Figure 1 demonstrates this interactive exploration. In particular, users can visualize any graph in the repository by simply selecting it from the left menu. 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.

2.4 Interactively Compare Graph Data

Graphs, nodes, and edges are easily compared across a wide range of important and fundamental graph statistics and parameters (e.g., max k-core number, total number of triangles, degree, max clique size, etc.). We note that subgraphs, nodes, and edges are easily compared to others within the same graph (e.g., nodes from a single network such as Twitter) or across multiple graphs, e.g., nodes from one network can be compared to nodes in another network. Figure 4 provides one example of how graphs can be interactively compared and gives intuition for the types of

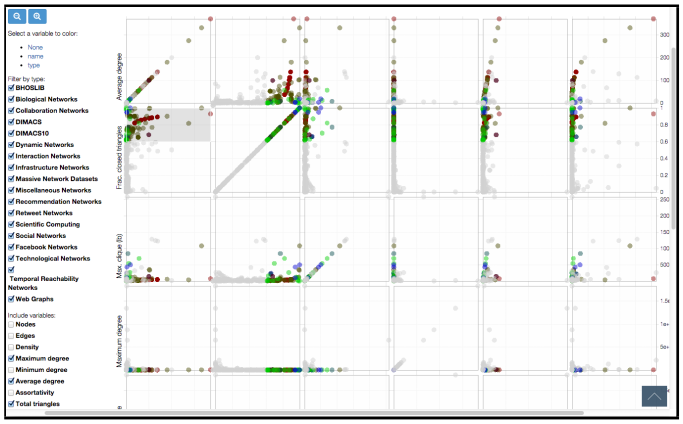


Figure 4: Interactively comparing nodes and graphs across a wide range of fundamental graph and node-level properties. Each data point represents a graph and each unique color represents the graph collection (e.g., social networks). In this example, we filter all graphs that have a global clustering coefficient (κ) less 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 $\kappa \geq 0.6$.

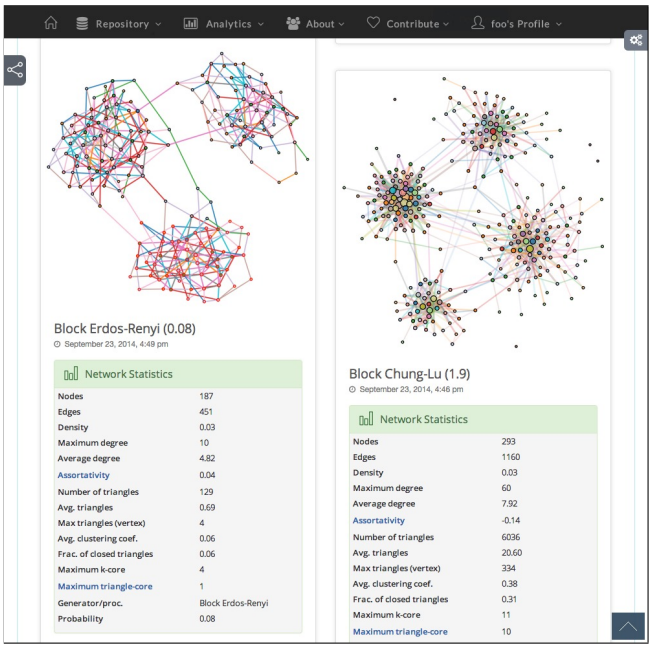


Figure 5: Users can create a profile and leverage many additional features, e.g., previous graph queries are automatically saved, as well as visualizations, and graph data from any shared or generated graphs. Additionally, graph visualization preferences can be saved and used throughout the site. The figure above shows a user's profile as well as two of their graph datasets that were uploaded and/or generated using NR's platform.

queries and/or questions that are possible. In particular, Figure 4 filters via any user selected constraint(s) and then highlights all such graphs (or nodes/edges) that satisfy it across all other interactive plots. We also allow the user to zoom-in and out on the plots interactively. Thus, in the case that graphs (data points) are overlapping and difficult to see when zoomed-out, the user can simply zoom into the data until a sufficient resolution is reached such that the graphs represents as points are no longer overlapping and the distinction between such graphs (and their statistics) are clearly observed. Graphs and/or collections of graphs can be filtered and/or colored using a variety of properties. Further, each graph statistic (variable) can be explored in greater detail by drilling down on a single statistic or set of statistics. For instance, the user may select one or more graph statistics for drilling. The first drilled statistic becomes the x-axis for all columns (representing graph statistics and/or features), and each column representing a graph statistic consists of the data points that match specific values for the drilled variables (beyond the first).

2.5 Community-oriented data repository

Another key advantage is in the community and collaborative nature of NR, e.g., users have the ability to collaboratively share data, knowledge, and insights. For instance, each dataset page has a place where users can share findings, visualizations, comments, as well as recently published articles that used that specific dataset. While other repositories are static and lack the ability for discussion and/or col-

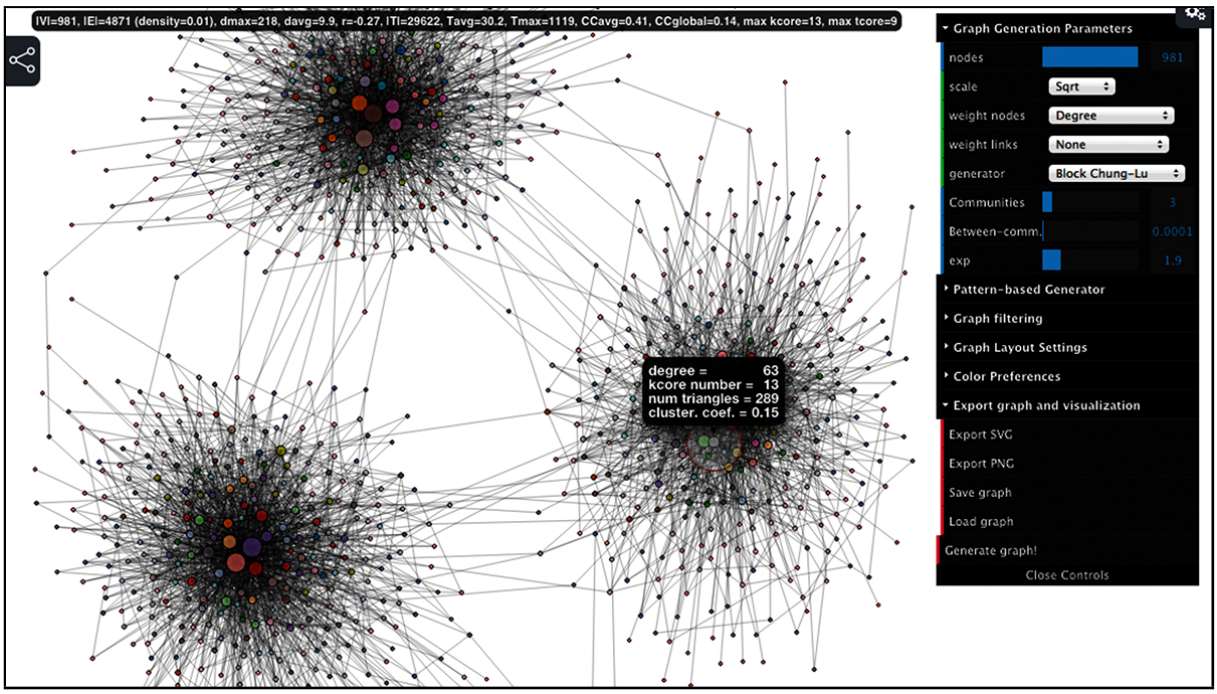


Figure 7: ***Interactive graph generation*** **using the proposed Block Chung-Lu graph model to capture community-structure.**

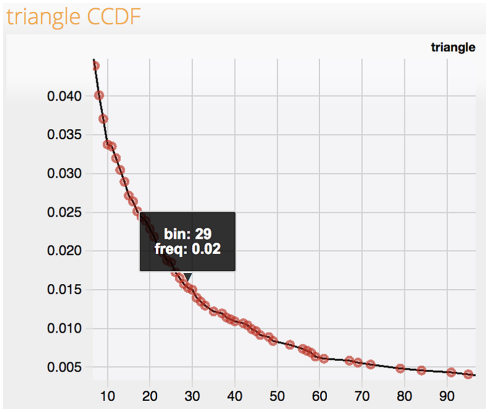


Figure 6: Interactive plot of the triangle count CCDF

laboration, we hope that a community/collaborative repository helps researchers to define which datasets are useful as benchmarks for specific tasks (e.g., anomaly detection, prediction, community detection, etc.).

2.6 Metadata and Documentation

We have tried to ensure that all graphs in the repository are sufficiently referenced and documented. However, users may also provide additional documentation, observations, post corrections, add recent papers that utilize a given dataset, etc (see Section 2.5). Thus, the documentation and our understanding of the graph data will improve over time by users and contributors that help refine the documentation to make it easier for others to understand.

2.7 Considerations for Big Graph Data

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, see [1]. These techniques are extremely effective for sampling node features and visualizing the structure and connectivity of the graphs. As an aside, the fast and scalable community and role discovery methods used in Figure 3 are also useful for visualizing big graph data.

2.8 Graph Computations

At the heart of the interactive platform lies the high-performance parallel graph analytics engine. It 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 when compared on the overlapping computations (e.g., triangle counting).

2.9 Interactive Graph Generators

NR provides a flexible interactive tool for synthetic graph generation and visualization. The platform has a variety of options for generating synthetic graphs of a particular size. For instance, NR allows for the generation of graph data using standard model-based graph generators, such as Erdos-Renyi (ER) graph model [4], Barabasi-Albert (BA) graph model [3], and Chung-Lu (CL) graph model [2]. In addition, we also provide pattern-based generators where users can generate graphs by connecting various graph patterns, such as nodes, edges, cliques, stars, cycles, and chains of various sizes. Furthermore, the platform provides a hybrid graph

generator that allows users to import various subgraph patterns into their synthetic graphs generated from any number of model-based generators (e.g., Chung-Lu, Erdos-Renyi, etc). We also provide various features to save graphs (e.g., as an edge list) as well as export images of visualizations (e.g., svg, png) generated by the user (both to disk and to their corresponding NR user account/profile).

3. CONCLUSION

This paper proposed the first graph data repository with a visual interactive analytics platform for real-time exploratory analysis and visualization of graph data. The platform was shown to be useful for interactively exploring, visualizing, and comparing graphs. While current data repositories are static, non-collaborative, and focus on downloads, NR is dynamic, community/collaborative-oriented, and focused on providing visual interactive analytics as well as making data easily accessible and understandable. Further, the platform was also shown to be flexible, easy-to-use, and fast for exploring and discovering valuable insights into the data. Overall, we believe the main contribution of this work is the proposal of an interactive data repository and demonstration of its utility and ability for facilitating scientific research. Finally, this work also serves as a basis for incorporating visual interactive analytics in current and future data repositories.

Acknowledgments

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

4. 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] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

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

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