

The Network Data Repository with Interactive Graph Analytics and Visualization

Ryan A. Rossi and Nesreen K. Ahmed

Dept. of Computer Science
Purdue University
West Lafayette, IN 47906

Abstract

(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, NR incorporates many social and collaborative aspects that facilitate scientific research, e.g., users can discuss each graph, post observations, and visualizations.

Discussion

This paper presents (NR) — the first *interactive data repository* with real-time graph analytics and visualization. NR has hundreds of graphs and network datasets for users to download (and share). However, the key factor that differentiates NR from other repositories (SNAP; UCI ML Repository) 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. NR 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

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

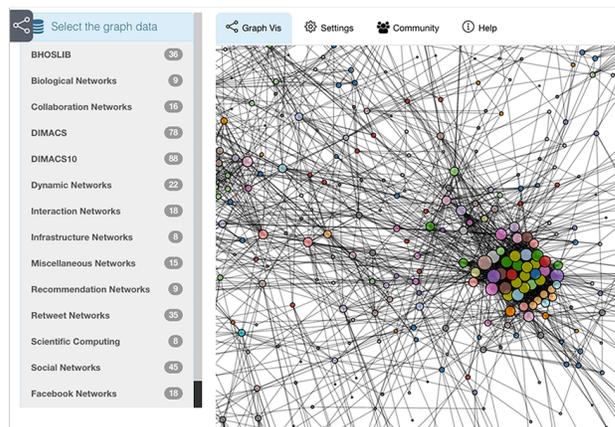


Figure 1: Visualize graph structure and discover valuable insights using our interactive graph visualization platform. Compare with hundreds of other networks across many different collections and types.

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 Repository, SNAP), 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. The goal of NR is to make it easy for users to discover key insights into the data quickly with little effort, while also providing a medium for researchers to share data, visualizations, and insights. In addition, we also make it easy for users to upload and visualize their own data.

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 flexibility to interactively plot and visualize the data according to the properties and characteristics of interest to them. Researchers can begin analyzing and investi-

gating 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.

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. 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 in 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 real-time.

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/subgraphs through interactive comparisons across a wide range of graph properties. We also provide many other interactive analysis tools, e.g., *interactive graph clustering tasks* such as role discovery and community detection.

Big graph data may also be interactively explored and visualized using NR. For instance, users can interactively explore a range of distributions from a wide variety of important graph properties and statistics. We also leverage 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 (Ahmed, Neville, and Kompella 2014).

Despite the increasing interest in graph data and algo-

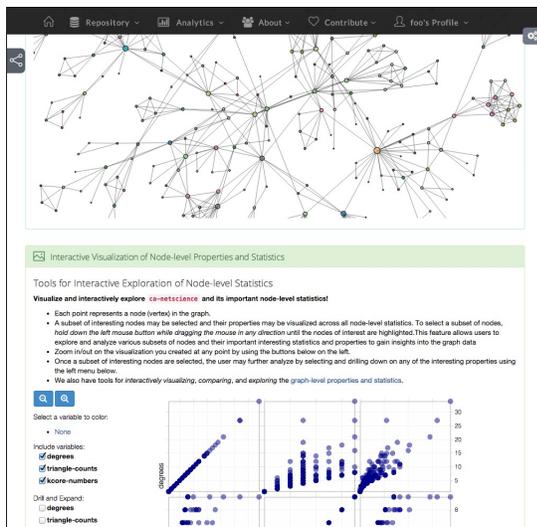


Figure 2: A snapshot of a graph's page showing the interactive graph structure visualization and node-level statistics for ca-netscience graph. Note that each graph is automatically processed and assigned a unique URL for reference purposes making it easy for others to obtain the exact data for experimental and evaluation purposes. The page also contains other analytics based on graph-level/point statistics, node and edge-level statistics and distributions.

rithms, there still remains a lack of standard benchmark datasets for many problems and research areas. Unfortunately, most research uses proprietary data and/or some pre-processed 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. 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. In addition, NR is a community-oriented repository that allows users to discuss, share observations, recent findings/papers, and any other insights. This would facilitate comparisons of various algorithms and models. For more details, we refer the reader to a longer version of this paper (Rossi and Ahmed 2014). We summarize a few of the contributions and features below.

- An interactive data repository where researchers can compare, explore, search, and analyze over 500+ graphs interactively in real-time via NR's web-based platform.
- Interactive visualization and exploration of the graph structure and connectivity patterns (e.g., nodes and edges).
- Global network statistics and parameters (e.g., triangle counts, max clique size, 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).
- Interactive visualizations and plots of key statistical distributions of each network (e.g., degree distribution).
- Community-oriented data repository where users can create profiles donate datasets, share visualizations, and insights, as well as save their synthetically generated networks and visualizations created using NR.
- Interactive graph generation including model-based (e.g., Chung-Lu (Aiello, Chung, and Lu 2001), BA (Barabási and Albert 1999), pattern-based (e.g., cliques, stars, cycles, etc.), and hybrid generators that combine both.

References

- Ahmed, N. K.; Neville, J.; and Kompella, R. 2014. Network sampling: From static to streaming graphs. *TKDD* 8(2):7:1–7:56.
- Aiello, W.; Chung, F.; and Lu, L. 2001. A random graph model for power law graphs. *Experimental Mathematics* 10(1):53–66.
- Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *science* 286(5439):509–512.
- Rossi, R. A., and Ahmed, N. K. 2014. Networkrepository: A graph data repository with visual interactive analytics. *arXiv:1410.3560*.
- SNAP. <http://snap.stanford.edu/data/index.html>.
- UCI ML Repository. <http://archive.ics.uci.edu/ml>.