

net.science: A Cyberinfrastructure for Sustained Innovation in Network Science and Engineering

Nesreen K. Ahmed², Richard A. Alo³, Catherine T. Amelink⁴, Young Yun Baek¹, Aashish Chaudhary⁵, Kristy Collins⁴, Albert C. Esterline⁶, Edward A. Fox⁴, Geoffrey C. Fox⁷, Aric Hagberg⁸, Ron Kenyon¹, Chris J. Kuhlman¹, Jure Leskovec⁹, Dustin Machi¹, Madhav V. Marathe¹, Natarajan Meghanathan¹⁰, Yasuo Miyazaki⁴, Judy Qiu⁷, Naren Ramakrishnan⁴, S. S. Ravi¹, Ryan A. Rossi¹¹, Rok Sosic⁹, Gregor von Laszewski⁷

¹University of Virginia, VA, USA, ²Intel Labs, CA, USA, ³Florida A&M University, FL, USA,

⁴Virginia Tech, VA, USA, ⁵Kitware, NY, USA, ⁶North Carolina A&T State University, NC, USA,

⁷Indiana University, IN, USA, ⁸Los Alamos National Laboratory, NM, USA, ⁹Stanford University, CA, USA,

¹⁰Jackson State University, MS, USA, ¹¹Adobe Research, CA, USA

Abstract—Networks have entered the mainstream lexicon over the last ten years. This coincides with the pervasive use of networks in a host of disciplines of interest to industry and academia, including biology, neurology, genomics, psychology, social sciences, economics, and cyber-physical systems and infrastructure. Several dozen journals and conferences regularly contain articles related to networks. Yet, there are no general-purpose cyberinfrastructures (CI) that can be used across these varied disciplines and domains. Furthermore, while there are scientific gateways that include some network science capabilities for particular domains (e.g., biochemistry, genetics), there are no general-purpose network-based scientific gateways. In this work, we introduce *net.science*, a CI for Network Science and Engineering, that is designed to be a community resource. This paper provides an overview of *net.science*, addressing key requirements and concepts, CI components, types of applications supported, and various dimensions of our evaluation process.

Index Terms—cyberinfrastructure, network science, *net.science*, network science gateway

I. INTRODUCTION

A. Background

Networks are pervasive in society. Today's urban infrastructures are networked and include power grids, communication networks, and transportation networks. There are also a variety of social networks that form the fabric of various kinds of diffusion processes, such as scholarship, commerce, epidemics, and fads. The need to understand these and other networks has led to a growing science of networks, and a multi-disciplinary community of researchers engaged in their study. Beyond the fact that network science has emerged as a discipline in its own right, it has also enabled fundamental discoveries in other scientific disciplines [5], [6].

A Google search of *networks* returns about 4.8 billion hits and Google Scholar returns over 5 million entries. Companies such as Facebook, Twitter, Amazon, Yahoo, Google, LinkedIn and Akamai use networks as a central concept in their businesses. Multiple reports by National Academies, and popular

journals such as Science, Nature, and PNAS contain regular articles and special issues devoted to this topic. More than 20 journals and 15 international conferences are devoted to findings about networks, or have them as a significant theme. Many books have been written on this topic. They span many areas including biology, social sciences, information sciences, health sciences, infrastructure systems, business, economics, communication networks, cybersecurity, mathematics, environmental sciences, and ecology.

B. Motivation for a CI for Network Science

The foregoing makes it clear that network science finds use in a wide array of disciplines. Furthermore, it is often the case that a particular network concept has wide applicability across domains. For example, computing connected components of a graph will have different semantics in different disciplines. But, from a computational standpoint, the algorithm and its execution are the same across applications; analysts, domain experts, and users apply their own semantics to these results. Similarly, there are several anomaly detection algorithms that represent systems as networks and that are used in domains as diverse as transportation, water distribution, crime detection, image analysis, computer security, event detection, and genomics (e.g., [3]). These considerations indicate that a cyberinfrastructure (CI) for network science would have broad applicability and therefore high value.

The XSEDE science gateway¹ lists more than 40 gateways, but none is devoted to network science in general. The Science Gateways Community Institute (SGCI)² has roughly 600 gateway entries. A search of the latter for *networks* and *network science* returns over 15 individual gateways for network visualization and analyses for genomics, proteins, metabolics, biology, genes and transcription factors, biochemistry, systems biology, computer networks, environmental and earth

¹<https://www.xsede.org/ecosystem/science-gateways>

²<https://catalog.sciencegateways.org/#/home>

sciences, diseases and health informatics, and computational neuroscience. Hence, current gateways do not fill the wide-ranging needs described above.

C. net.science: A CyberInfrastructure for Sustained Innovation in Network Science and Engineering

The National Science Foundation (NSF) has recently funded the development of a CI for network science and engineering, called *net.science*. This CI will contain applications (i.e., software tasks) that compute on, e.g., analyze and visualize, (principally) network data and produce networks from data, among other features. It will be largely domain-agnostic, although users can supply domain-specific applications (see below). Further, its services will be openly accessible, consistent with other scientific gateways.

Our goal of making *net.science* a *community resource* is predicated on achieving the requirements specified in Section II. The *net.science* CI must be well-regarded so that content creators feel that they derive value from making their products available through this CI. Similarly, the value must be clear to users. Hence, there must be a symbiotic relationship among contributors, users, and infrastructure; see Figure 1. Our view is consistent with NSF's vision and requirements for cyberinfrastructures [2].

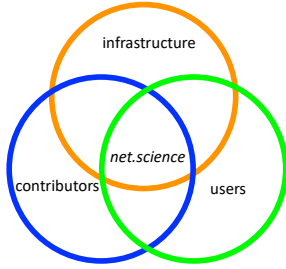


Fig. 1. *net.science* as a symbiotic cyberinfrastructure, bringing together producers (contributors of software, (network) data, and learning materials) and consumers (users).

The purpose of this paper is to provide an overview of the CI and its evaluation. It is early in the project, but we will have periodic releases with new features. More information is provided on the web site: <http://net.science>.

II. net.science SYSTEM

Several aspects of the CI are presented here. Our approach is to use existing middleware and infrastructure components to the extent possible. As one example, various workflow engines and languages are being evaluated as part of Section II-E.

A. Architecture

Figure 2 provides a system view of *net.science*. Compute resources on which the system will run are given in aqua (near the bottom). There is a set of infrastructure management services such as system monitoring and logging, and configuration and resource management. A workflow engine is shown in orange, enabling seamless integration of data, software tools, libraries, tasks, and web applications (web apps). Of the illustrative services within the Workflow Engine,

the Deep/Machine Learning service is expanded to the right, providing illustrative libraries with which tasks are built. Above these boxes is the application programming interface (API) layer, with services and a digital library (DL) with data stores. The application layer, at the top of the graphic, is exposed through the API layer, as are the workflow engine and data.

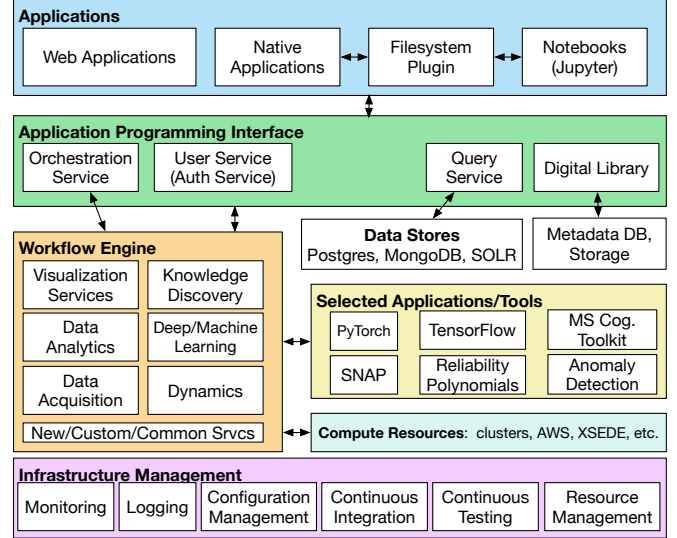


Fig. 2. Layered system view of *net.science*. See text for a description of these layers.

B. APIs (Application Programming Interfaces)

The APIs in Figure 2 will be programming language (PL) independent using HTTP. The current plan is to wrap this HTTP-native API with each of Python and Java to produce PL-specific APIs in these commonly used languages, for ease of use by *net.science* users. (Languages such as C++ will use the native API.) A broader view of the *net.science* API is provided below in Section II-H.

C. Common Services

Common services cut across particular applications and are of general use. Examples are Figure 2's Workflow Engine boxes. For instance, a library containing graph structural analysis methods would be used to construct a set of software tasks, but visualization of results would be accomplished with a more general visualization service. Dynamics means services to compute dynamical processes on networks (e.g., the spread of contagion).

D. Illustrative Software Libraries and Tasks

Software libraries and frameworks such as NetworkX [8], SNAP [10] and Repast [13] are state-of-the-art and are used pervasively due to, among things, their wide ranges of capabilities. Serial and parallel libraries and tasks will be included in *net.science*. Table I summarizes a few state-of-the-art tasks that are each supported by different groups. The purpose of presenting these applications is that each is a more targeted type of application that *net.science* seeks to attract owing to their novelty.

TABLE I

ILLUSTRATIVE LIST OF SELECTED STATE-OF-THE-ART APPLICATIONS TO INCLUDE IN *net.science*.

Item	Application (Task)	References
1	<i>Reliability polynomials.</i> Provide ways to assess resilience of networks.	[7]
2	<i>Anomaly detection.</i> Methods to identify anomalous subgraphs in network representations of various systems.	[3]
3	<i>SAT solvers.</i> Satisfiability problem solvers are useful in many applications (e.g., model checking, analyzing discrete dynamical systems).	[14]
4	<i>Knowledge graphs.</i> Methods to answer complex logical queries on large-scale incomplete knowledge bases.	[12]
5	<i>Software workbench for data mining tasks.</i> Tools and toolkits like Weka.	[17]

E. Workflows

User-composable workflows, comprised of several applications, are another key to the success of *net.science* [4]. In workshops on the use of the CINET CI [1], users were adamant that they wanted “one-stop shopping” software tools. For example, they did not want to export results from a network structural analysis software (task) into some other tool for subsequent analysis. They wanted all functionality in one place, so that analysis processes can be completed within the same system. Operations such as data transformation of output, for input to a down-stream process, will be handled seamlessly through datatypes that are used in pre-defined transformations.

F. Network Data

The system will make available a wide range of networks from many domains (e.g., social science, genomics, biology, and social and online media). Graphs may be directed and undirected, and labeled, and *net.science* will initially contain networks with hundreds of millions of nodes and edges. Graph sizes will increase over time. There will also be collections of graphs; for example, networks from the same population whose nodes and edges vary over discrete time snapshots.

G. FAIR Principles

FAIR data principles (Findability, Accessibility, Interoperability, and Reusability) [18] will be followed in maintaining data. For example, a file service and a DL provide expandable sets of searchable metadata per digital object (DO) type.

H. APIs and X-As-A-Service

Figure 3 is a logical view of types of interactions that we envision with *net.science*. Users may request services through a user interface (UI) that processes information through the *net.science* API. Users may also write scripts and workflows to perform single jobs or compositions of them. One example is the use of (Jupyter) notebooks to provide an interactive and adaptive analysis approach. Third-party applications such as R-Studio or web applications such as those from the Network Repository [11] will be able to perform operations on data stored in *net.science*. Genomics data may be retrieved from the PATRIC software system [16], and used to construct networks. Coordinating data across tasks within workflows

(data engineering), for example, is a technical challenge [4], [15], represented by the big data system within *net.science*. As an example, for tasks that require different network formats, transformations (either direct or through a base format) will be executed. A goal is to expose *net.science* as a gateway.

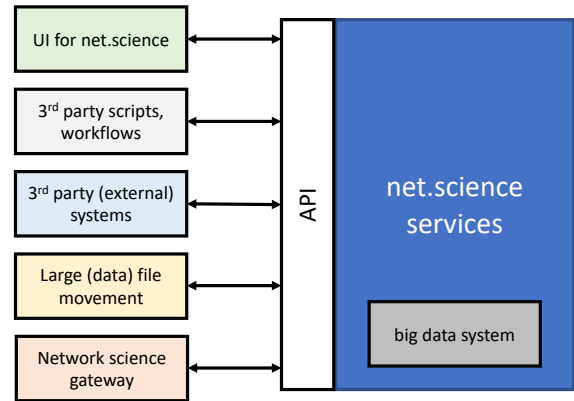


Fig. 3. Logical view of external interactions of users, software, and other systems with *net.science* through the API.

I. Requirements for net.science

Selected system requirements are provided below. Several of these requirements are common to many systems. User contributions and automatic attribution of their contributions is different from some CIs.

- 1) *Security and permissions.* Users must have credentials to access the system and particular resources. Users that contribute data and software (i.e., codes) will control their accessibility (e.g., private, exposed to a subset of users, public), and these contributors can change accessibility over time.
- 2) *Provenance.* Applicable to data (including software), the steps and inputs leading to each DO’s creation must be documented and retrievable.
- 3) *Reproducibility.* All job specifications (inputs, outputs, code) must be stored so that analyses may be rerun (including by others possessing suitable permissions).
- 4) *Extensibility, innovation.* It must be straightforward for content producers to add their software and data to *net.science*.
- 5) *Rigor, reliability.* Procedures must ensure that contributed content is checked for correctness.
- 6) *Continuous testing and integration.* These are required to ensure proper functioning of the system.
- 7) *User-base contributions.* The system must be sufficiently well-regarded to motivate content creators (code, unique data sets, learning materials) to contribute to the system. This may include parallel versions or otherwise improved versions (in terms of performance or capabilities) of current tasks. This does not preclude authors from also hosting these same products on their own websites.
- 8) *Attribution.* Providers (producers) of content must be given full and prominent credit for their contributions. They must be able to search the system to obtain metrics

on the use of their supplied materials. They must feel, even if maintaining their software and data through their own web sites and outlets like GitHub, that they benefit from including their DOs in *net.science*.

- 9) *Adaptability, Stack Overflow-like content.* Users must be able to comment on system resources (e.g., to request help from other users) and also to request new capabilities. This is also crucial for adaptability: users must be able to specify their needs so that producers—from the community at-large—can respond to them.
- 10) *Target user base.* Our goal is to attract individuals and groups from industry, applied R&D and government laboratories, educators, college and university students (including undergraduates) across academic disciplines, and high school students.
- 11) *Metrics.* Metrics must be kept on use of codes and data so that contributors can receive credit and usage information and understand how contributions are used.
- 12) *Target platforms.* *net.science* is intended to run on commodity and high-performance computing (HPC) clusters, including GPUs and Hadoop clusters [19], and cloud resources.

It should be noted that there are caveats to almost all of these requirements. For example, particular data may have been generated with software that was later found to have contained a bug, and the reworked software may have different inputs. This affects requirement 3.

III. SYSTEM EVALUATION

Figure 4 gives an overview of our system evaluations for *net.science*. We will evaluate along three broad dimensions: scientific merit (row 1), educational value (row 2), and outreach value (row 3). For each dimension, activities, assessments, and outputs are listed. Each of these will be judged by a panel of academic experts. Scientific evaluation (row 1) includes system and task performance evaluations. The educational component (row 2) will include classroom use of *net.science* and evaluations of students, implemented in a coordinated effort with the professors that are teaching the network science-related courses. Outreach (row 3) will include workshops and demos, to solicit feedback. Based on two previous workshops on a prior cyberinfrastructure [1], we anticipate participants from a range of academic disciplines (e.g., computer science, neuroscience, liberal arts, statistics, bioinformatics, and demography) and roles (e.g., researchers, college students, high school students). These activities are also consistent with recommended practices to promote CI adoption [9].

Acknowledgment. We thank the anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation (NSF) Grant Nos. OAC-1916805, OAC-1835631, OAC-1835598, DGE-1545362, and IIS-1633363.

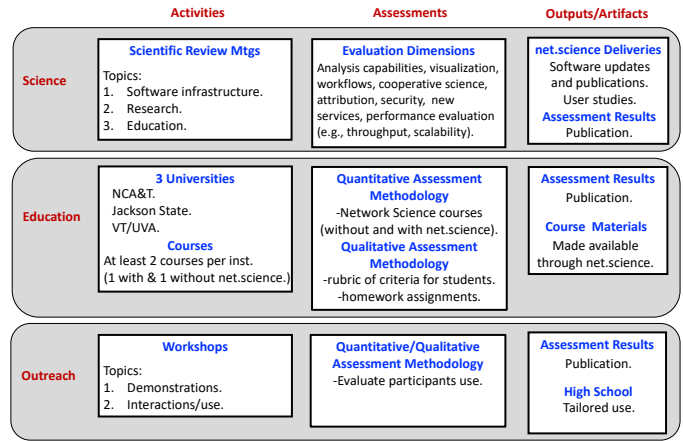


Fig. 4. Dimensions scientific, educational, and outreach assessment of *net.science*.

REFERENCES

- [1] S. Abdelhamid, M. Alam, R. Alo *et al.*, “CINET 2.0: A cyberinfrastructure for network science,” in *e-Science*, vol. 1, 2014, pp. 324–331.
- [2] Anonymous, NSF, “Transforming science through cyberinfrastructure: NSF’s blueprint for a national cyberinfrastructure ecosystem for science and engineering in the 21st century,” National Science Foundation, Tech. Rep., 2019, <http://ci-coe.sci.utah.edu/images/resources/nsf-aci-blueprint.pdf>.
- [3] J. Cadena, F. Chen, and A. Vullikanti, “Graph anomaly detection based on Steiner connectivity and density,” *Proceedings of the IEEE*, vol. 106, no. 5, pp. 829–845, 2018.
- [4] D. A. (Chair), “Realizing opportunities for advanced and automated workflows in scientific research: Second meeting,” *The National Academies of Sciences, Engineering, Medicine*, 2020, <https://www.nationalacademies.org/event/03-16-2020/realizing-opportunities-for-advanced-and-automated-workflows-in-scientific-research-second-meeting>.
- [5] Committee on Network Science for Future Army Application, National Research Council, *Network Science*. National Academies Press, 2005.
- [6] K. Coronges, A. Barabasi, and A. Vespignani, *Future Directions of Network Science: A Workshop Report on the Emerging Science of Networks*. Blacksburg, VA: Virginia Tech Applied Research Corporation, 2016.
- [7] S. Eubank, M. Youssef, and Y. Khorramzadeh, “Using the network reliability polynomial to characterize and design networks,” *Journal of Complex Networks*, vol. 2, no. 4, pp. 356–372, 2014.
- [8] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX,” in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 2008, pp. 11–15.
- [9] B. Le, F. Escalera, K. Jitkajornwanich, and K. F. Kee, “External communication to diffuse science gateways and cyberinfrastructure as innovations for research with big data,” in *Gateways*, 2019.
- [10] J. Leskovec and R. Sosič, “SNAP: A general-purpose network analysis and graph-mining library,” *ACM Trans. Intell. Syst. Technol.*, 2016.
- [11] “Network Repository: An interactive scientific network data repository,” <http://networkrepository.com/>, 2020.
- [12] H. Ren, W. Hu, and J. Leskovec, “Query2box: Reasoning over knowledge graphs in vector space using box embeddings,” in *ICLR*, 2020.
- [13] “The Repast Suite,” <https://repast.github.io>, 2020.
- [14] “Information regarding SAT solvers,” www.satlive.org, 2018.
- [15] D. Sculley, G. Holt, D. Golovin *et al.*, “Hidden technical debt in machine learning systems,” in *NIPS*, 2015.
- [16] A. R. Wattam, J. J. Davis, R. Assaf *et al.*, “Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center,” *Nucleic Acids Res.*, vol. 45, pp. D535–D542, 2017.
- [17] “Weka: The workbench for machine learning,” <https://www.cs.waikato.ac.nz/ml/weka/>, 2020.
- [18] M. Wilkinson, M. Dumontier, I. Aalbersberg *et al.*, “The FAIR guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, pp. 160018–1–160018–9, 2016.
- [19] B. Zhang, Y. Ruan, and J. Qiu, “Harp: Collective communication on Hadoop,” in *2015 IEEE International Conference on Cloud Engineering*, 2015, pp. 228–233.