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

Nesreen Ahmed[2], Richard Alo[3], Catherine Amelink[4], Young Yun Baek[1], Aashish Chudhary[5], Kristy Collins[4], Albert Esterline[6], Edward Fox[4], Geoffrey Fox[7], Aric Hagberg[8], Ron Kenyon[1], Chris J. Kuhlman[1], Jure Leskovec[9], Dustin Machi[1], Madhav V. Marathe[1], Nataragan Meghanathan[10], Yasuo Miyasaki[4], Judy Qiu[7], Naren Ramakrishnan[4], S. S. Ravi[1], Ryan Rossi[11], Roc Sosic[9], Gregor von Laszewski[7]

[1]*University of Virginia, VA, USA,* [2]*Intel Labs, CA, USA,* [3]*Florida A&M University, FL, USA,*
[4]*Virginia Tech, VA, USA,* [5]*Kitware, NY, USA,* [6]*North Carolina A&T State University, NC, USA,*
[7]*Indiana University, IN, USA,* [8]*Los Alamos National Laboratory, NM, USA,* [9]*Stanford University, CA, USA,*
[10]*Jackson State University,* [11]*Adobe 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, psychology, 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 Engineering and Science, that is designed to be a community resource. This paper provides an overview of *net.science*, addressing key requirements and concepts, CI components, the types of applications that our CI will support, and various dimensions of our evaluation process.**

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

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 is 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 [6], [7].

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., [4]). These considerations indicate that a cyberinfrastructure (CI) for network science would have broad applicability and therefore high value.

The XSEDE science gateway[1] 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 sciences, diseases and health informatics, and computational

The CINES project is supported by NSF Grant OAC-1916805.

[1]https://www.xsede.org/ecosystem/science-gateways
[2]https://catalog.sciencegateways.org/#/home

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 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). Furthermore, 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 [3].

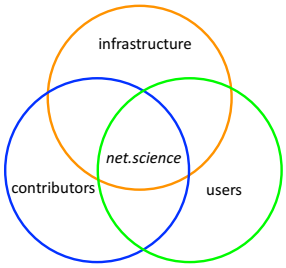

Fig. 1. *net.science* as a symbiotic cyberinfrastructure, bringing together producers (contributors of software, (network) data, and learning materials) and consumers (users).

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.

A. Architecture

Figure 2 provides a system view of *net.science*. Compute resources on which the system will run are given in blue (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. It is a key component from a usability viewpoint, since this service enables users to compose tasks to complete larger units of work. This is a primary value proposition of the system: use of tools within the system (as opposed to outside of *net.science*) enables seamless integration of data, software tools, libraries, applications, and web applications (web apps). Of the illustrative services within the Workflow Engine, the Deep/Mach. Learning service is expanded to the right, providing illustrative applications. Above these boxes is the 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.

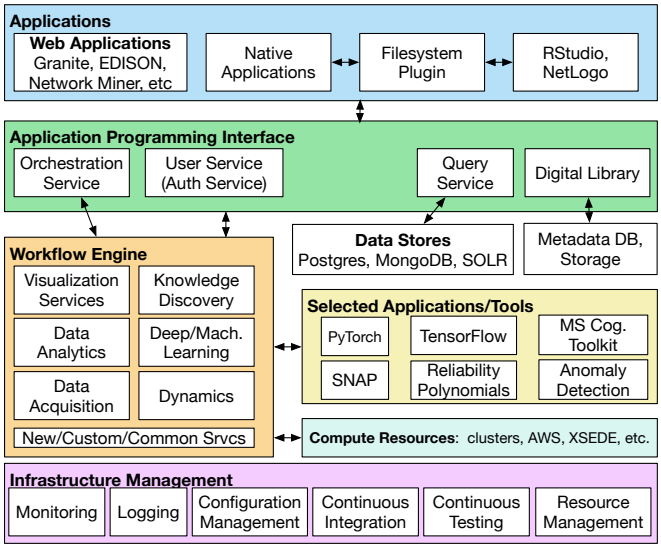

Fig. 2. Layered system view of *net.science*. See text for a description of these layers.

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 classified as a particular application, 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 Applications

The value of libraries and applications is measured by their use and the types of problems they solve. Software libraries and frameworks such as NetworkX [1], SNAP [15], and Repast [12] are state-of-the-art and are used pervasively due to, among things, their wide ranges of capabilities. Serial and parallel libraries and applications will be included in *net.science*. Table I summarizes a few state-of-the-art applications that are each produced 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	References
1	*Reliability polynomials.* Provide ways to assess resilience of networks.	[8]
2	*Anomaly detection.* Methods to identify anomalous subgraphs in network representations of various systems.	[4]
3	*SAT solvers.* Satisfiability problem solvers are useful in many applications (e.g., model checking, analyzing discrete dynamical systems).	[13]
4	*Knowledge graphs.* Methods to answer complex logical queries on large-scale incomplete knowledge bases.	[11]
5	*Software workbench for data mining tasks.* e.g., Weka.	[17]

E. Workflows

User-composable workflows, comprised of several applications, are another key to the success of *net.science* [5]. In workshops on the use of the CINET [2] CI, 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 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. Also, this is a major opportunity for analysis-time collaboration among multiple users.

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 of 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. A file service and a digital library provide expandable sets of searchable metadata per digital object 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. Third-party applications, such as R-Studio or web applications such as those from the Network Repository [10] 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 [5], [14], represented by the big data system within *net.science*. A goal is to expose *net.science* as a gateway.

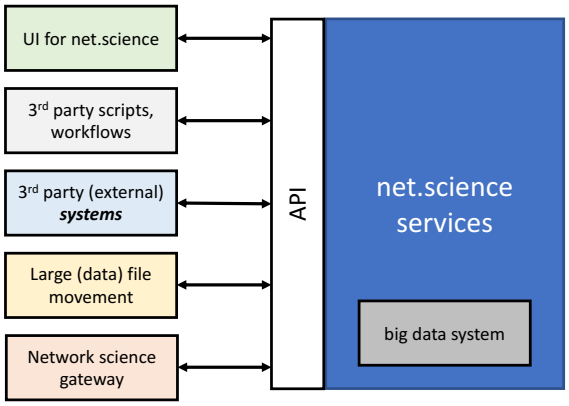

Fig. 3. Logical view of external interactions of users, software, and other systems with *net.science* through the API.

I. Requirements for net.science

With particular aspects of the system described above, selected system requirements are provided below.

1) *Security and permissions.* Users must have credentials to access the system and particular resources. An illustrative use case is that a researcher may be generating data for a publication and may not want to make that data public until after a manuscript has been accepted for publication.
2) *Provenance.* Applicable to data (including software), the steps and inputs leading to each digital object'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 straight-forward 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 creators of content (code, unique data sets, learning materials) to contribute to the system. This does not preclude authors from also hosting these same products on their own websites. Also, contributions may be for particular application domains; we do not preclude this. These contributions are the key to sustained innovation and maintaining the system as state-of-the-art.
8) *Attribution* for providers of software, data, and other digital objects (DOs) such as learning materials. 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* (e.g., for integration in workflows with other data and tools).

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. (Currently, AI and machine learning are hot topics. But others will surely arise, and *net.science* must be responsive to changes in user emphases. A Stack Overflow-like way to request content will assist this adaptability.)
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, to understand how their 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 (science, row 1), educational value (row 2), and outreach value (row 3). For each dimension, a related set of activities is listed. For example, scientific merit will be judged in three categories: software infrastructure, research capabilities, and education. Each of these will be judged by a panel of academic experts using the listed evaluation dimensions under assessment. 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. For example, assignments completed with and without *net.science* will be used to assess the differential in learning provided by the CI. Outreach (row 3) will include workshops and demos, to solicit feedback. Based on two previous workshops on a prior cyberinfrastructure [2], 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].

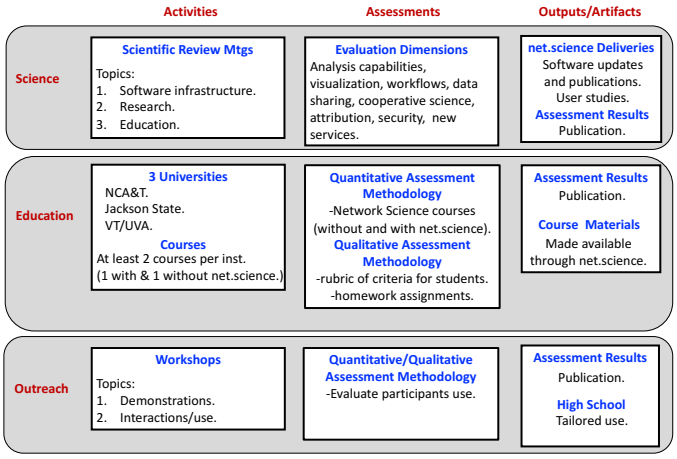

Fig. 4. Various dimensions scientific, educational, and outreach assessment of *net.science*.

References

[1] "NetworkX," https://networkx.github.io/, 2020.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] Committee on Network Science for Future Army Application, National Research Council, *Network Science*. National Academies Press, 2005.
[7] 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.
[8] 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.
[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] "Network Repository: An interactive scientific network data repository," http://networkrepository.com/, 2020.
[11] H. Ren, W. Hu, and J. Leskovec, "Query2box: Reasoning over knowledge graphs in vector space using box embeddings," in *ICLR*, 2020.
[12] "The Repast Suite," https://repast.github.io, 2020.
[13] "Information regarding SAT solvers," www.satlive.org, 2018.
[14] D. Sculley, G. Holt, D. Golovin *et al.*, "Hidden technical debt in machine learning systems," in *NIPS*, 2015.
[15] "Stanford Network Analysis Project," http://snap.stanford.edu, 2020.
[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. 160 018–1–160 018–9, 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.