PROJECT SUMMARY

In today's world, human behavior, social networks, and civil infrastructures are closely intertwined. Decisions made by humans impact physical infrastructures (e.g., power and road networks) and the behavior of infrastructure under loads imposed by these decisions alters how humans interact with physical infrastructure systems. The need to represent and analyze such socially coupled interdependent infrastructure networks to assess short- and long-term policy questions poses significant scientific and technical challenges. Recent quantitative advances in computational science, network science, artificial intelligence and behavioral sciences have created new opportunities for analyzing and studying complex, nonlinear, coupled behavior of such interdependent infrastructures.

The specific goals of the proposal are: (i) Mathematical techniques and associated algorithms for generating ensembles of infrastructure networks and social interactions; (ii) Mathematical techniques for measuring the structural and dynamic properties of such networks, and (iii) Methods for validation and verification of these interdependent synthetic populations and network ensembles.

Intellectual Merit: The novel aspects of our proposal include the following. (i) A focus on very large realistic socially-coupled infrastructure networks – the primary focus here will be coupled power and communication networks. The proposed work will result in open-access data sets for the larger academic community that are unique, extensive and realistic. (ii) Development of highly scalable data-driven dynamic methods that allow us to develop methods for sensitivity analysis (SA), uncertainty quantification (UQ) and verification and validation (V&V). The research will provide new ways to characterize and assess vulnerabilities inherent in today's societal infrastructures. One unique feature of the proposed research is to study vulnerabilities resulting from interdependencies between infrastructures as well as vulnerabilities and risks that are human-mediated; that is, vulnerabilities resulting from interaction of individuals with the infrastructure as well with each other.

Broader Impact: The proposal will bridge the gap between the social, computing and engineering sciences, so as to effectively address fundamental questions in the science of complex networks. The tools and techniques developed under the project will also be of interest to the broader community of analysts and researchers interested in complex network science.

Beyond the stated objectives, the project will advance the state of the art in social science, urban planning as well as develop new methods in theoretical computer science, electrical engineering, motivated by applications in social science, critical infrastructure protection and urban planning, enhancing the support of decision making in complex situations. This includes: (i) novel, efficient, distributed, exact, and provably approximate algorithms for assessing structural and dynamic properties of other large complex networks (e.g., transportation networks, and financial networks); (ii) general methods that can be used to assess SA, UQ and V&V for networked systems. The effort will advance graduate, undergraduate and professional curricula. Data sets and methods produced as a part of project will be made available to scientists.

1 Introduction and Objectives

PPD-21 lists 16 critical infrastructures [34]. These infrastructures have multiple interdependencies and failures can easily cascade between them. These systems have three common characteristics: the first is overloading of the remaining system when elements are removed from service; the second is that these systems are inherently robust until the point of collapse; and the third is that these systems have protection systems. Many of the incorrect operations of the protection/control system can be characterized as hidden failures because they remain hidden until there are disturbances near the protection equipment. The 1965, 1977 and most recently 2003 blackout all involved protection equipment not necessarily as the initiating event but as a contributor to the cascading nature of the event. Understanding these couplings and cascading effects is central to successful planning and development of resilient next-generation infrastructures.

In order to study coupled infrastructures, a central challenge is the lack of realistic data sets that have detailed representations of each network as well as the interactions across infrastructures and the interactions with the human population of operators, emergency responders, and consumers. Harmonizing and integrating representations of subsets of infrastructures is a challenge that all CRISP type projects will face. High quality data sets and a standardized representation will provide a platform for accelerating research into these domains as well as the possibility of comparing outcomes across research projects.

Human mediated interdependencies. Physical interdependencies between infrastructures are now well accepted. In contrast, human mediated dependencies that arise through the role of expert operators, emergency responders, and consumers is not well understood. As a simple example, we can combine estimates for how much electricity a person uses at home or at work together with time dependent estimates of the number of people in each activity and the geographic distribution of home and work locations to arrive at a time-dependent geographic distribution of demand for electricity. The required estimates can be calibrated and validated against demand data from typical days as well as entirely different kinds of data, such as traffic counts, land use, and time use surveys. But whereas the data from typical days offers no insight into abnormal days, by capturing the behavioral causes of observed aggregate demand at the individual level, we can indeed generalize to hypothetical situations.

1.1 Proposed effort

In our work spanning 20+ years, we have developed methods for synthesizing representations of populations, social interactions, and networks, using over a dozen public and commercial datasets. Specifically, transport, power, and communication systems (networks and human aspects) were synthesized. One shortcoming of our and others' related work is that it has focused on single instances of such infrastructures and social systems. This has made assessing important questions related to verification and validation (V&V), uncertainty quantification (UQ) and sensitivity analysis (SA) harder to address.

Here, we propose to develop ensembles of such coupled systems, and methods to assess UQ and V&V issues. These ensembles of coupled power and communication networks will provide realistic data for algorithm development, UQ, and SA. We will develop mathematically rigorous techniques and associated algorithms, focusing on developing scalable methods that can produce a large number of ensembles. The proposed effort will consist of three distinct topics:

Topic 1: *Mathematical techniques and associated algorithms for generating ensembles of infrastructure networks and social interactions.* The starting point will be single instances, using our prior methods. We will develop techniques for local structure preserving perturbations to generate the ensembles, understanding the coupled systems, with physical as well as human mediated inter-dependencies. We will develop simple data management tools for storing, accessing and querying these network ensembles. The key aspect we will focus on is the development of a standard representation of critical infrastructures and their interactions, and querying properties across such ensembles.

Topic 2: *Mathematical techniques for measuring the structural and dynamic properties of inter-dependent power and communication networks.* Many open source codes exist which we will use when appropriate. These measures are needed to ensure that the ensemble networks satisfy basic structural and dynamical properties. For example, we might want to ensure that the contagion properties of the networks within an ensemble are preserved. To do this, we will use "contagion test signals" based on published mathematical models. As appropriate, our goal is to use known and published codes for this to avoid duplication of effort.

Topic 3: *Methods for validation and verification (V&V) of these interdependent synthetic populations and infrastructure network ensembles.* Illustrative questions to be studied include: (i) comparison of two coupled networks to determine their relative quality, (ii) study of the uncertainty of dynamical outcomes over ensembles of coupled networks. Methods developed in Topic 2 will be critical here, providing methods to test network quality and form the basis for V&V and UQ. The project will strive to use methods and data sets already produced and available to the research community. We will also actively work with other community partners to obtain relevant data sets; in this regard, for example, we will work with the NSF South Big Data Hub.

Outcomes. The primary outcomes are: (i) generation of ensembles of synthetic communication and power networks for Washington, DC—these will incorporate physical and human-mediated inter-dependencies, with basic methodology useful for other coupled networks; (ii) generation of ensembles of synthetic populations and micro-activities pertaining to the use of these two networks and facilitating their coupling; and (iii) a methodology for V&V for assessing the qualities of such interdependent synthetic systems. These ensembles will be available for use by the broader scientific community, made available by leveraging our prior NSF funded project CINET (a cyber-infrastructure for network science). CINET allows web-based access to data sets as well as methods for structural and dynamical analysis. We will use CINET for wide dissemination of methods and data.

How will this be useful: The proposed effort will yield data sets useful in understanding the secondary and tertiary effects of large-scale disruption of power and communication networks. Additionally, the methods utilized will allow analysts to study important questions related to SA, UQ and V&V for such systems. For instance, one could study how certain algorithms for detecting a cascading power failure, and its impact on communications, depend on the interacting communication and power networks.

How these data can contribute to better understanding of demand for ICI-based services: The proposed data and methods can be used for planning new smart grids, understanding the impact of large scale disruptions on these networks and assessment of response strategies.

An illustrative example. A concrete example of this is to study the effects of a hypothetical air burst of a large nuclear device that can cripple the two networks (in isolation) but also together as cascades formed when the protective devices fail due to high altitude electro-magnetic pulse (HEMP). These protective devices are networked for developing a detailed situation assessment in case of local disturbances. HEMP is likely to make these networks ineffective. The failure of the power network causes further loss of power to communication systems that has direct economic and social impact. The impact is highly dependent on the scenario, and computing the impact will require a detailed understanding of the functioning social systems. The scenario is important and pertinent in today's geo-political state of affairs. Earlier work by ORNL and others suggested that the networks would not be impacted significantly. What has changed (since that study in 1984) is the role of communication networks in control and situation assessment. Additionally, electrical and communication networks that are realistic are not available anymore due to proprietary and security reasons. No such study exists, yet it is extremely important from the standpoint of protection and response of our ICIs. Understanding the cascading failures will require a detailed sensitivity analysis and multiple scenario representations. The specific study is beyond the scope of the project – we believe that the datasets we develop can assist undertaking such a study.

2 Preliminary and related work

Power & Communication networks. We use a first-principles based approach for modeling mobile social networks and spectrum demand [20]; this is in contrast with random network based models, which are commonly used in literature, e.g., [27]. Our approach integrates over a dozen public and commercial data sets with various social and economic theories, and produces a synthetic yet realistic spatial, dynamic, relational network of individual agents (devices, individual people, base stations) at a highly resolved spatial scale of a few meters, and a temporal scale of a few seconds, with links representing various relationships, including, spatial collocation, sessions and their details (length, type, etc.). This tool was used within the NPS-1 study for modeling the communication infrastructure and user call behavior in the Washington D.C. region [30].

In the past, constituent elements of the electrical network were available for research. This has changed radically in the recent years due to security, privacy and proprietary concerns. To overcome this challenge, the proposed research will create a realistic representation of the actual grid with the help of publicly available data sources and expert knowledge [11]. The structure of the network has a profound impact on the dynamical process e.g. its ability to transfer current, its vulnerability and reliability in the event of failures and therefore it is important to have at least a somewhat realistic electrical network to work with. Zussmann and collaborators have investigated statistical models for geographically correlated failures in power and communication networks [3]. The goal of their work was determination of likely vulnerable locations in communication networks under single and multiple attacks.

Coupled Systems. Over the last 20+ years, we have developed an integrated suite of analytic tools and simulation models to analyze the effects on a collection of social and physical infrastructures after a WMD attack [2, 11, 13, 30, 45, 48, 52, 66]. The collection of models was developed in the context of a low yield nuclear detonation taking place in Washington DC, and the framework incorporates the infrastructures addressed by this proposal.

What is missing. There has been a lot of interest in extending results in individual networks to cascading failures in such interdependent infrastructures, e.g., [9, 10, 24, 25, 36, 44, 46, 49, 60, 68, 71, 73]. Buldyrev et al. [25] have studied models of cascading failures in interdependent random networks, and derive analytic bounds on the critical fraction of failed nodes. Leicht et al. [24, 44] consider a slightly more general model, and find that the percolation threshold in individual networks can be significantly lowered when the interdependencies are considered. The work of Brummitt et al. in [24] is of similar nature. Cascading effects in *scale-free* networks have been investigated by Asztalos et al. [9].

While these results give valuable insights into network vulnerability, their practical implications are limited as real networks often cannot be captured adequately by idealized random graph models. The shortcomings in these modeling approaches and their implications for use in policy have been pointed out by domain experts in several articles and reports, e.g., Alderson and Doyle [6].

The work described above will provide an excellent starting point for the proposed work. This includes basic networks and behavioral representations of the populations. Many components of the network were based on proprietary data sets – methods will need to be developed to develop general guiding principles from these data sets. For example, detailed location of several base stations is obtained from commercial vendors. We will derive general rules for where such base stations are located and then use these rules to synthesize the wireless network.

3 Proposed Work.

3.1 Topic 1: Algorithms for generating ensembles of coupled networks

Network	Scale	Datasets and description
Power network: transmission	$\approx 10^3 - 10^4$ nodes (e.g., substations, generators, control devices and sensors)	GIS imagery, PSS/E, historical data from utilities and expert knowledge. Networks have typically low treewidth [23].
Power network: distribution	$\approx 10^5 - 10^6$ nodes (e.g., homes and work locations)	Spatial layout of road network and built infrastructure, utility data, expert knowledge. Typically trees, especially closest to the demand locations.
Communication infrastructure	$\approx 10^2 - 10^3$ nodes (e.g., cell towers, switching centers)	Cell tower data from TowerMaps, queries to the APIs from Sprint, AT&T and the FCC Antenna Registration System, expert knowledge. Spatial network
Mobile call graph	$\approx 10^6 - 10^7$ nodes, representing call data records	Constructed from an activity based synthetic population model, using data on cell phone ownership and call behavior data. Networks typically have high clustering and low diameter.

Table 1: Summary of the scale and datasets used for the infrastructure network representations.

1. Generating Basic physical networks. We discuss methods for generating physical networks and the human elements that couple these networks.

Power Networks. We will develop methods and underlying theory that will yield "first principles based models" of a regional power network using generative processes. These models will yield representations of power networks that are statistically similar to real power networks. This estimated, synthetic electrical grid will be built for the Northeastern US using open source information such as, expert knowledge, data from Google Earth and the Power System Simulator Software (PSS/E), as summarized in Table 1. We will consider open data from diverse sources, e.g., the Library of Congress archives provide addresses

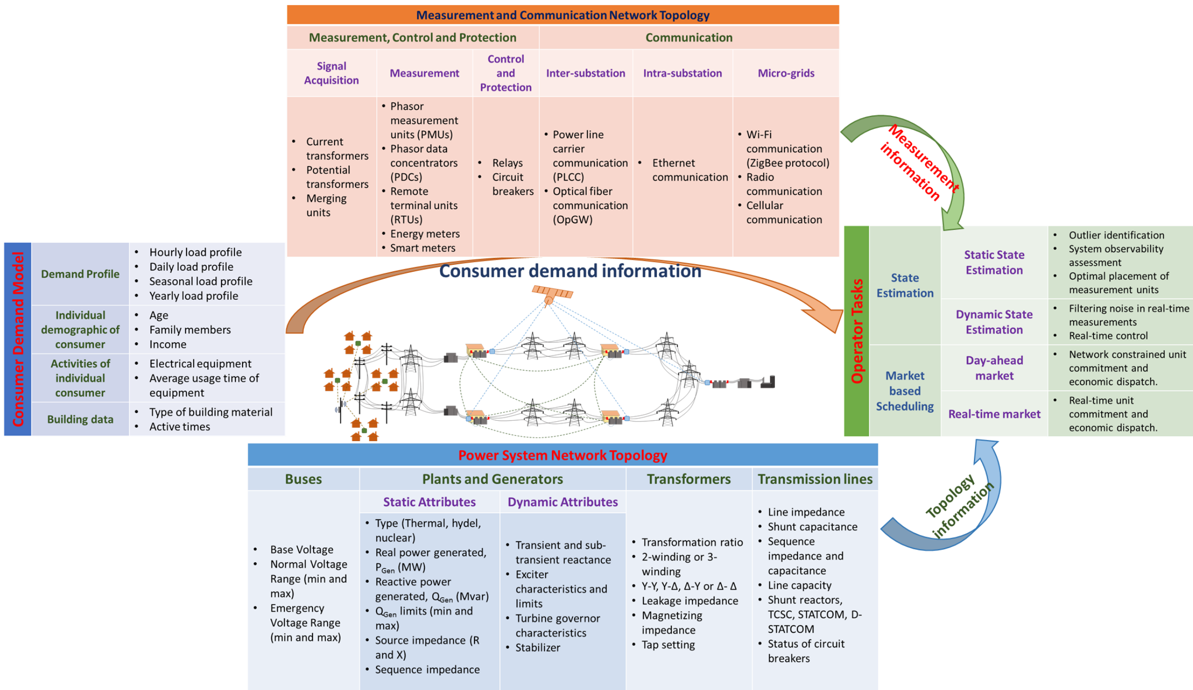

Figure 1: Illustration of the power network components. The figure illustrates four basic components: the physical network and the associated control and communication network make the power system; the consumers and operators makeup the human-mediated interactions between the networks. A similar figure for communication networks is omitted due to lack of space.

of several substations; many transmission manuals provide configuration of generating stations, line connections, voltage levels, and equipment codes for the high voltage lines; and reliability plans by electric companies give raw information on underground and overhead lines. A PSS/E model of the power system can be used as a guideline for determining the interconnectivity and network topology of the grid; relative distances available through transmission line parameters can be used as estimates.

Communication networks. We propose to develop a first-principles based approach for modeling mobile social networks and spectrum demand [19]; this is in contrast with commonly used random network based models, e.g., [27]. Our approach integrates over a dozen public and commercial data sets with various social and economic theories, and produces a synthetic yet realistic spatial, dynamic, relational network of individual agents (devices, individual people, base stations) at a highly resolved spatial scale of a few meters, and a temporal scale of a few seconds, with links representing various relationships, including, spatial collocation, sessions and their details (length, type, etc.). Most cellular infrastructure data is proprietary, and is difficult to obtain. We will construct a partial representation of the cellular infrastructure of a region by integrating data summarized in Table 1. We will then use methods developed in [19] for generating spatio-temporal call volume.

2. Incorporating physical and human-mediated coupling. Physical coupling of the networks is obtained by embedding these networks in space and using the embedding to model the dependencies. Simple rules such as proximity models to more complicated models that capture hidden interdependencices will be studied. Specifically, we will consider a multi-level network representation, in which the infrastructures are different levels, with edges representing the coupling between components, e.g., an edge (u, v) from a substation u to a cell tower v, to which it supplies power.

Human-mediated coupling will be obtained by incorporating human behaviors and their interactions with the power and communication infrastructures into the generated dynamic data. We will use a representation that allows modeling temporally extended behaviors [55]. This is done by using a semi-Markov decision process (SMDP) representation, where high-level behaviors, denoted options, are modeled as policies over low-level actions [62]. Each behavioral option has triggering and termination conditions, defined in terms of the agent and environment states. For example, a behavior such as “looking for family members” can be triggered by a condition where an emergency occurs (such as a natural disaster) and the agent is not with its family members. It can be terminated by finding the family members or being able to establish communication with them and learning that they are safe. The policy, or option, defining this behavior would involve the agent moving towards its family members and attempting to call them on the phone. Thus, the status of the power and/or communication infrastructure can play an important role in determining human behavior. In turn, behavior can affect infrastructures. For example, mass evacuation can cause traffic jams and overloading of the communication infrastructure. Similarly, behaviors can be modeled for infrastructure operators based on schemas they follow for normal control of the infrastructure as well as for mitigating potential failure conditions.

3. Generating network ensembles with structure preserving perturbations. We will use a broad class of constrained edge perturbations for generating ensembles, referred to as "edge shuffles" in labeled graphs. The basic edge shuffle operation is to rewire a pair of randomly selected edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ by (u_1, v_2) and (u_2, v_1); it is easy to see that this operation preserves the degrees of all the nodes. In general, nodes are associated labels, e.g., degree and demographic properties. We extend this to assortative edge shuffles, in which the edges e_1 and e_2 are rewired only if node properties such as degree and age distribution are preserved [41, 53, 54, 57]. However, they might completely disrupt the multi-scale structure in the network, so *we propose a new form of constrained edge shuffles, which preserve a given set of constraints*, such as: (i) network connectivity within and across given subsets, (ii) constraints on labels and functions of those, for specific nodes or sets of nodes. An example of the former is a constraint (i.e., upper and lower bounds) on the size of a cut $\delta(S,T) = |\{e = (u, v) : u \in S, v \in T\}|$. Cut constraints generalize degree sequence and assortativity constraints, which are commonly studied in network science [35, 41, 53, 54]. Such edge shuffles produce a distribution on graphs which satisfy the specified constraints. Such Markov processes are used in sampling from Exponential Random Graph Models (ERGM) [58, 61] to generate random graphs with specific properties. However, our focus is not on *generating* random graphs, but rather on creating controlled, local perturbations of a given graph.

Extension of these methods to produce ensembles of coupled networks is challenging. These ensembles will have to maintain the physical coupling constraints (e.g. substation at location (x, y) should not supply power to communication equipment beyond a certain range), as well as human-mediated coupling constraints (e.g. demands on power and communication networks needs to to obey basic temporal constraints). The dependencies are often captured using simple logical formulas, e.g. (If substation A goes down then communication gear in neighborhood Y will be affected). Probabilistic rules are often used in the real world and can be perturbed to obtain the needed ensembles.

4. Representations and storage of networks, coupling and human interactions. The lack of standardized formats for these aspects is a challenge to CRISP work. Naturally, the breadth of ICIs, as well as their rich, diverse and domain-specific features makes such a task highly complex. We will start tackling this in the context of power and communication networks with human operators while keeping the larger ICI picture in mind. For basic analyses such as connectivity or redundancy, it may be sufficient to know individual network structure, and possibly cross-edges between networks. For a power flow analysis, one may need physical characteristics to support a DC or an AC model. In the case of designing approaches employing communication-based control systems to power network cascading failures, details of SCADA-type systems will matter.

In our proposed work we will leverage our existing knowledge of digital library technologies [1] and ICIs [12, 30, 45, 55] to design representations of ICIs and human interactions that permits: (i) capturing basic network properties, (ii) permits domain-specific data layers to be added to each network as extensions, (iii) supports representation of interactions between networks (basic and extensions), and (iv) representations capturing mechanisms of human interaction with ICIs. This standardization work is required for all serious CRISP efforts, and is a natural step forward in accommodating serious progress.

3.2 Topic 2: Measuring structural & dynamic properties of coupled networks

Our approach for precisely describing and analyzing ICIs is based on *graph dynamical systems (GDS)*, a formal framework the team has developed over the last 20 years [14–17, 43, 47, 50, 51]. GDS were structured precisely to capture the dynamics of networked infrastructure systems with human interaction. Important design criteria of the GDS framework include strong support for mathematical, computational, algorithmic and statistical analysis, as well as naturally generating computational models that map efficiently to current high performance computing platforms.

For computer simulation models of ICIs one will frequently synthetize network(s), add suitable models to be run over the network, and then measure selected aspects of the dynamics. Clearly, one can measure graph theoretic properties, dynamical properties, as well as measures over the dynamical properties. Which measures are relevant will depend on context.

We will develop a set of measures for assessing coupled power and communication networks. The measure will include both graph-theoretic quantities and dynamic quantities, the latter category being measures associated to dynamic processes taking place over the networks. Such processes will include power/communication-specific measures, but will also include abstract measures such contagion processes. Contagion processes are lighter weight processes that, for example, give insight on diffusion and cascade dynamics over networks [22, 26, 29]. Particular emphasis will be placed on measures that assess properties arising through coupling of infrastructure networks and through human interaction.

The solution will be implemented using a flexible and extensible plugin design capturing provenance and computation specific details. This solution will be available in a standalone fashion, but will also be exposed in parts through CINET [1, 31].

3.3 Topic 3: V&V, SA and UQ for the coupled networks

We will undertake a detailed statistical analysis of the algorithms and the coupled networks they produce. We will go beyond predictive and postdictive validity – structural validity is important in our context. Validation based on just observed data is incapable of capturing the entire structural range —- only those modes that happened in the real world appear in the measured data. Further, there is usually insufficient information for postdiction, causing the estimated models fitting well only for relatively sparse data.

1. Global and cross sensitivity analysis of coupled networks with perturbation. We propose to conduct the global sensitivity analysis for the entire system, and develop cross sensitivity analysis between the coupled networks in the system. As the observed data is often limited to some extent, we will adopt the constraint perturbation (e.g., constraint edge shuffle) to enable more modes occurred in the constructed networks. Then the global sensitivity analysis will be conducted to evaluate the dynamic qualities of the system, such as cascade dynamics over networks. The functional ANOVA models [39] will be used for such a global analysis, exploring the effect of the individual network on the entire system. Moreover, we will develop cross sensitivity analysis to understand how the performance of one network will be affected by the change of other coupled networks. The additive Gaussian process [33] will be decouple the uncertainty of one network in terms of other coupled networks. By linking the coupled networks in a unified framework, it would better borrow strength from the relatively sparse data for model validation. Also the developed additive Gaussian process will also be useful for model calibration [40].

2. The joint effect of network and local dynamics on system dynamics. Using our earlier work, we will build a computational theory for understanding sensitivity of system dynamics to network structure and local functions. We propose to develop sparse experimental design techniques to study this systematically for coupled telecommunication and power networks. Such a design strategy is to allocate design points to be space-filling for the coupled networks systems, while maintain good stratification for each individual network. Thus it is advantageous to study both system and local dynamics. The modern response surface methodology [18, 42] will be used to investigate the sensitivity for different system dynamics, as a function of different parameters. Extending prior work, we will also study stability of cascade dynamics, identifying network structure conditions and function properties under which the systems are robust.

3. Statistical designs supporting efficient SA and UQ for coupled network. As noted, parametrization of the demographic/agent related attributes and of the coupled network dynamics will allow us to address the challenges of scalability for models of cascading failures. Because of the complexity of coupled networks, the input parameters will easily contain a large number of continuous, discrete, and nested variables. It is not possible to exhaust all possible parameter combinations for sensitivity analysis (SA) and uncertainty quantification (UQ). We will overcome this by using adaptive designs; this will greatly enhance the efficiency of getting into the important region of the coupled network systems.

First, the marginally-coupled design (MCD) [32], containing both multi-levels and continuous settings, will be used for such construction. MCD ensures that the design points of continuous inputs forming a Latin hypercube design obey space-filling property [59]. Moreover, when projecting data points to each level of any discrete input (e.g., type of each network in the coupled networks system), the corresponding design points for continuous inputs also form a small Latin hypercube design with space-filing property. This makes SA and UQ more accurate, while keeping the computational time reasonable. Next, we will sequentially generate a small set of design points based on the latest estimated model, taking the network coupling into consideration in form of constraints. The expected-improvement (EI) criterion [67] will be used to select next small set of design points. The use of EI criterion can ensure that the selected points will improve the prediction accuracy of key quantities. We will further consider multi-level sparse designs as well as novel designs based on formulating the design problems as location theoretical problems (dispersion problems). The proposed designs will extend the Latin hypercube design [56] and orthogonal arrays [37]. Such proposed designs achieve maximum stratification in one and two dimensions with appropriate number of well chosen design points that are significantly smaller than a full factorial design. This approach enables the joint study of the main effect of each parameter and the nonlinear interaction of any two parameters under k levels while keeping time complexity low.

4 Team, Management, Evaluation and Education & Outreach

Team qualifications. The team has a proven track record in this area, having developed and delivered a number of large-scale and highly used environments. Many of the co-PIs and senior personnel have extensive experience in program leadership on large, interdisciplinary programs, including in the role of principal investigator. Our team has been working on aspects related to the topic for over 20 years. We were the founding members of the DHS NISAC program while at the Los Alamos National Laboratory, and developed the urban infrastructure suite (UIS). This was the first time coupled representations of realistic infrastructures where created and where human behaviors were incorporated. Ongoing work with the DoD will be leveraged for the proposed work. Our recent tutorials on the subject provide further information, see [63].

Team Member Roles. This project involves a multi-disciplinary group from multiple academic and research organizations. Team members have collaborated closely for many years. This includes large projects for several federal agencies, including DoD, DHS, IARPA, NIH, and CDC, many of which involved different kinds of studies on network science.

- Marathe (Fellow: ACM, IEEE and AAAS) – *Expertise:* Computer science, communication networks, and critical infrastructure. *Role*: (Tasks 1, 2 & 3).
- Deng – *Expertise:* Statistical science, uncertainty quantification, sensitivity analysis, Bayesian inference, data science, and experimental design. *Role:* (Tasks 1 and 3)
- Mortveit – *Expertise:* Mathematics, modeling, design and data management. *Role:* (Tasks 1 & 2).
- Phadke (Inventor of PMUs; Franklin medal winner; Member NAE) – *Expertise:* Power Systems, Network security, Wide area measurements. *Role:* Tasks 1 and 2.
- Swarup – *Expertise:* AI, Data Science, Machine Learning and Synthetic Systems. *Role:* (Task 1 & 3) Responsible overseeing the human aspects of the network synthesis process
- Vullikanti (DOE and NSF early career award recipient for work on related topics) – *Expertise:* Communication networks, Algorithmics and Network Science. *Role:* Tasks 1 & 2.

Project management and coordination. We will have regular WebEx meetings every two weeks. With the exception of Prof. Phadke, who will visit every three months, the team is located at Blacksburg. We will actively forge relationships with DHS COE and the DHS NISAC program to explore possible ways to share data and methods. A mid-project meeting in DC will be organized at VT-NCR to solicit input from experts in the field.

Coordination Mechanisms. We will take active steps to make our tools and data sets available to the larger academic community. The VT team and its collaborators have developed CINET, a cyber-infrastructure middleware for network science, as a part of another NSF-funded project. It is a pervasive computing environment for researchers and educators, and provides a large collection of networked data sets and support for running efficient algorithms for networks and dynamics on them. We plan to leverage this effort by using CINET and the NSF South DataHub facilities (VT is a member) to disseminate selected synthesized infrastructure networks and to expose several data sets that we have synthesized and analyzed pertaining to realistic social networks.

Evaluation. We will work with domain experts to evaluate the quality of the synthetic data. Phadke has extensive contacts with the power companies. Marathe and Vullikanti are also associated with WirelessVT that provides access to a number of communication network companies. Open source data sets available

will be used for validation. The data sets will be made accessible to the larger community via standardized formats; this is an important consideration for effective use of such coupled networks. While developing standards is out of scope for the present proposal, the work will begin a dialogue on this important subject.
Education, Outreach and Synergistic Activities 2-3 graduate students will be hired under the project. Dr. Swarup was a co-organizer of the NSF BIGDATA program and Big Data Hubs & Spokes PI meeting this year and will be a co-organizer again next year. Webinars will be organized to announce the data sets created as a part of the project as well as get input from scientific community to improve the quality of the data sets. We have a close working relationship with DHS NISAC, other DHS COEs and DTRA. All these organizations are intimately involved in developing technologies related to the proposed work. Whenever possible we will leverage the ongoing work in these organization.

5 Expected Project Significance

Intellectual Merit. The novel aspects of our proposal include the following. (i) A focus on very large realistic socially-coupled infrastructure networks – the primary focus here will coupled power and communication networks. The proposed work will result in open-access data sets for the larger academic community that are unique, extensive and real-world. (ii) Development of highly scalable data-driven dynamic methods that allow us to develop methods for sensitivity analysis, uncertainty quantification and verification and validation. The research will provide new ways to characterize and assess vulnerabilities inherent in today's societal infrastructures. One unique feature of the proposed research is to study vulnerabilities resulting from physical and human-mediated interdependencies.

Broader Impact. The proposal will bridge the gap between the social, computing and engineering sciences, so as to effectively address fundamental questions in the science of complex networks. The tools and techniques will also be of interest to analysts and researchers interested in complex network science. Beyond the stated objectives, the project will advance the state of the art in social science, urban planning as well as new methods in theoretical computer science, electrical engineering, motivated by applications in social science, critical infrastructure protection and urban planning and the support of decision making in complex situations. This includes: (i) novel, efficient, distributed, exact, and provably approximate algorithms for assessing structural and dynamic properties of other large complex networks (e.g. transportation networks, financial networks, etc); (ii) general methods that can be used to assess SA, UQ and V&V for networked systems. The effort will advance graduate, undergraduate and professional curricula. Data and methods produced as a part of project will be made available to scientists.

6 Results from Prior NSF Support

Marathe and **Vullikanti** were supported by "CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science" (ACI-1443054, $1,125,000, 10/01/2014-09/30/2019). *Intellectual merit*. The goal is to advance high performance data analytics with efficient parallel algorithms, especially for static and dynamic network problems. Papers resulting from this work include: [4, 5, 7, 8, 21], one of which [5] was nominated for the best paper award at the 2016 Supercomputing conference. *Broader impacts*. A number of applications and technologies are enabled by the methods developed in this project, including network science, social networks and epidemiology. These algorithms are being made available through a cyber-infrastructure called CINET [1]. The project has partially supported 2 post docs and three graduate students. **Deng** serves as the PI for the NSF project: CMMI-1233571, $123,192, Duration: 09/01/2012-08/31/2015, titled "Collaborative Research: A Statistics-Guided Framework for Synthesis and Characterization of Nanomaterials". *Intellectual merit*. Several novel quantitative methods for improving key experimentation and characterization in the nonmanufacturing process, e.g., a mixed variance component model for estimation of the elasticity of 1D nanomaterials, a statistics-enabled predictive method for high-precision characterizations of surface properties of nanomaterials. *Broader impacts.* Research results are presented at several national and international conferences [32, 38, 64, 69, 70, 72]. Results from this project were presented to 3M representatives and the NSF Head of Engineering, respectively. Research components have been integrated into the PI's teaching of data analytics and design of experiments.
Swarup (VT) is a Co-PI in "IBSS: Understanding social diffusion dynamics among networked cognitive systems" (SMA-1520359, $950,000, 09/01/15-02/28/19). The intellectual merit is to integrate network science and cognitive science in the study of the diffusion of information. The broader impact is to involve students and researchers in the social sciences in computational cognitive modeling.

References

[1] S. Abdelhamid, R. Alo, S. Arifuzzaman, et al. CINET: A cyberinfstructure for network science. In *Proc. of the 8th IEEE International Conference on eScience*, Oct. 2012.

[2] A. Adiga, H. S. Mortveit, and S. Wu. Route stability in large-scale transportation systems. In T. Ito, C. Jonker, M. Gini, and O. Shehory, editors, *The Workshop on Multiagent Interaction Networks (MAIN), held in conjunction with AAMAS 2013*, St. Paul, MN, USA, May 7 2013.

[3] P. K. Agarwal, A. Efrat, S. K. Ganjugunte, D. Hay, S. Sankararaman, and G. Zussman. The resilience of wdm networks to probabilistic geographical failures. *IEEE/ACM Transactions on Networking*, 21:1525–1538, 2013.

[4] M. Alam and M. Khan. Parallel algorithms for generating random networks with given degree sequences. *International Journal of Parallel Programming*, to appear. available online, doi: 10.1007/s10766-015-0389-y.

[5] M. Alam, M. Khan, A. Vullikanti, and M. Marathe. An efficient and scalable algorithmic method for generating large–scale random graphs. In *SuperComputing*, 2016.

[6] D. Alderson and J. Doyle. Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40:839–852, 2010.

[7] S. Arifuzzaman, M. Khan, and M. Marathe. A fast parallel algorithm for counting triangles in graphs using dynamic load balancing. In *Proceedings of the 2nd Workshop on High Performance Big Graph Data Management, Analysis, and Mining (BigGraph), in conjunction with the IEEE International Conference on Big Data (IEEE BigData)*, Oct. 2015.

[8] S. Arifuzzaman, M. Khan, and M. Marathe. A space-efficient parallel algorithm for counting exact triangles in massive networks. In *Proceedings of the 17th IEEE International Conference on High Performance Computing and Communications (HPCC)*, August 2015.

[9] A. Asztalos, S. Sreenivasan, B. Szymanski, and G. Korniss. Distributed flow optimization and cascading effects in weighted complex networks. *Eur. Phys. Journal B*, 85(288), 2012.

[10] J. Banerjee, A. Das, and A. Sen. A survey of interdependency models for critical infrastructure networks. *arXiv preprint arXiv:1702.05407*, 2017.

[11] C. Barrett, K. Bisset, S. Chandan, J. Chen, Y. Chungbaek, S. Eubank, Y. Evrenosoglu, B. Lewis, K. Lum, A. Marathe, M. Marathe, H. Mortveit, N. Parikh, A. Phadke, J. Reed, C. Rivers, S. Saha, P. Stretz, S. Swarup, J. Thorp, A. Vullikanti, and D. Xie. Planning and response in the aftermath of a large crisis: an agent-based informatics framework. In *Proceedings of the 2013 Winter Simulation Conference*, 2013.

[12] C. Barrett, K. Bisset, S. Chandan, J. Chen, Y. Chungbaek, S. Eubank, Y. Evrenosoğlu, B. Lewis, K. Lum, A. Marathe, M. Marathe, H. Mortveit, N. Parikh, A. Phadke, J. Reed, C. Rivers, S. Saha, P. Stretz, S. Swarup, J. Thorp, A. Vullikanti, and D. Xie. Planning and response in the aftermath of a large crisis: An agent-based informatics framework. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, editors, *Proceedings of the 2013 Winter Simulation Conference*, 2013.

[13] C. Barrett et al. Social, health and socio-technical effects of an IND in the National Capitol: Demonstration by expanding CNIMS to involve detailed interdependent behavioral properties in national planning scenario 1. *Technical Report*, 2013.

[14] C. Barrett, H. Hunt III, M. Marathe, S. Ravi, D. Rosenkrantz, R. Stearns, and M. Thakur. Computational complexity of analyzing the dynamic reliability of interdependent infrastructures. In *Proceedings of the Third International Conference on Critical Infrastructures, Arlington, Virginia*, 2006.

[15] C. L. Barrett, K. Bisset, S. Eubank, V. S. A. Kumar, M. V. Marathe, and H. S. Mortveit. *Modeling and Simulation of Large Biological, Information and Soci-Technical Systems: An Interaction-Based Approach*, volume 64 of *Proceedings of Symposia in Applied Mathematics*, pages 101–148. American Mathematical Society, 2007.

[16] C. L. Barrett, H. B. Hunt III, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. On Special Classes of Sequential Dynamical Systems. *Annals of Combinatorics*, 7:381–408, 2003.

[17] C. L. Barrett, H. B. Hunt III, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. Predecessor Existence Problems for Finite Discrete Dynamical Systems. *Theoretical Computer Science*, 386(1-2):3–37, October 2007.

[18] K. W. Bauer Jr, G. S. Parnell, and D. A. Meyers. Response surface methodology as a sensitivity analysis tool in decision analysis. *Journal of Multicriteria Decision Analysis*, 8(3):162, 1999.
[19] R. Beckman, K. Channakeshava, F. Huang, J. Kim, A. Marathe, M. Marathe, G. Pei, S. Saha, and A. K. S. Vullikanti. Integrated multi-network modeling environment for spectrum management. *IEEE Journal on Selected Areas in Communications*, 31(6):1158–1168, June 2013.
[20] R. Beckman, K. Channakeshava, F. Huang, J. Kim, A. Marathe, M. Marathe, S. Saha, G. Pei, and V. A. Kumar. Integrated multi-network modeling environment for spectrum management. *IEEE Journal on Selected Areas in Communication*, 31:1158–1168, 2013. Special issue on Network Science.
[21] H. Bhuiyan, J. Chen, M. Khan, and M. Marathe. Fast parallel algorithms for edge-switching to achieve a target visit rate in heterogeneous graphs. In *Proceedings of the International Conference on Parallel Processing (ICPP)*, 2014.
[22] K. Bisset, J. Chen, C. J. Kuhlman, V. S. A. Kumar, and M. V. Marathe. Interaction-based HPC modeling of social, biological, and economic contagions over large networks. In *Proceedings of the Winter Simulation Conference*, pages 2938–2952, 2011.
[23] H. Bodlaender. Treewidth: Algorithmic Techniques and Results. In *Proc. MFCS*, pages 29–36, 1997.
[24] C. D. Brummitt, R. M. D'Souza, and E. A. Leicht. Suppressing cascades of load in interdependent networks. *PNAS*, 109(12):E680–E689, 2012.
[25] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028, 2010.
[26] M. M. C. Kuhlman, V.S.A. Kumar et al. Diffusion and blocking of complex contagions in bi-threshold systems. In *Proc. of AAAI Fall Symposium on Complex Adaptive Systems*, 2011.
[27] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, pages 483–502, 2002.
[28] Creative commons. https://creativecommons.org/licenses/by/4.0/.
[29] D. Centola. Failure in Complex Social Networks. *J. Math. Sociology*, 33:64–68, 2009.
[30] S. Chandan, S. Saha, C. Barrett, S. Eubank, A. Marathe, M. Marathe, S. Swarup, and A. K. Vullikanti. Modeling the interactions between emergency communications and behavior in the aftermath of a disaster. In *The International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP)*, Washington DC, USA, April 2-5 2013.
[31] CINET. http://cinet.vbi.vt.edu/cinet_new/.
[32] X. Deng, Y. Hung, and C. D. Lin. Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica*, pages 1567–1581, 2015.
[33] X. Deng, C. D. Lin, K.-W. Liu, and R. Rowe. Additive gaussian process for computer models with qualitative and quantitative factors. *Technometrics*, (to appear), 2016.
[34] Critical infrastructure sectors. https://www.dhs.gov/critical-infrastructure-sectors.
[35] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
[36] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin. Networks formed from interdependent networks. *Nature physics*, 8(1):40–48, 2012.
[37] A. S. Hedayat, N. J. A. Sloane, and J. Stufken. *Orthogonal arrays: theory and applications*. Springer Science & Business Media, 2012.
[38] R. Jin and X. Deng. Ensemble modeling for data fusion in manufacturing process scale-up. *IIE Transactions*, 47(3):203–214, 2015.
[39] C. G. Kaufman, S. R. Sain, et al. Bayesian functional {ANOVA} modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, 2010.
[40] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
[41] M. Khan, M. Youssef, X. Deng, S. Eubank, and M. Marathe. Joint sensitivity analysis of disease model and contact networks on epidemic dynamics, 2015. NDSSL Technical Report, Virginia Tech.
[42] A. I. Khuri and S. Mukhopadhyay. Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):128–149, 2010.
[43] R. Laubenbacher, A. Jarrah, H. Mortveit, and S. Ravi. A Mathematical Formalism for Agent-Based

Modeling. In R. Meyers, editor, *Encyclopedia of Complexity and System Science*, pages 160–176. Springer, 2009.

[44] E. A. Leicht and R. M. D'Souza. Percolation on interacting networks. *arxiv:0907.0894*, 2009.

[45] B. Lewis, S. Swarup, K. Bisset, S. Eubank, M. Marathe, and C. Barrett. A simulation environment for the dynamic evaluation of disaster preparedness policies. *The Journal of Public Health Management and Practice*, 19:S42–S48, Sep/Oct 2013. PMCID: PMC3962069.

[46] T. Macaulay. *Critical infrastructure: understanding its component parts, vulnerabilities, operating risks, and interdependencies*. CRC Press, 2016.

[47] M. Macauley and H. S. Mortveit. Cycle equivalence of graph dynamical systems. *Nonlinearity*, 22(2):421–436, 2009. math.DS/0709.0291.

[48] M. Marathe et al. Computational architecture and an implementation of a synthetic information and modeling environment for reasoning about human-initiated crisis. *Tech. Rep.*, 2013.

[49] J. V. Milanovic and W. Zhu. Modelling of interconnected critical infrastructure systems using complex network theory. *IEEE Transactions on Smart Grid*, 2017.

[50] H. Mortveit and C. Reidys. Discrete sequential dynamical systems. *Discrete Mathematics*, 226:281–295, 2000.

[51] H. Mortveit and C. Reidys. *An Introduction to Sequential Dynamical Systems*. Universitext. Springer Verlag, 2007.

[52] National Security Staff, Interagency Policy Subcoordination committee for preparedness and response to Radiological and Nuclear threats. Planning guidance for response to a nuclear detonation. 2010.

[53] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(208701), 2002.

[54] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 286(026126), 2003.

[55] N. Parikh, S. Swarup, P. E. Stretz, C. M. Rivers, B. L. Lewis, M. V. Marathe, S. G. Eubank, C. L. Barrett, K. Lum, and Y. Chungbaek. Modeling human behavior in the aftermath of a hypothetical improvised nuclear detonation. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Saint Paul, MN, USA, May 2013.

[56] J.-S. Park. Optimal latin-hypercube designs for computer experiments. *Journal of statistical planning and inference*, 39(1):95–111, 1994.

[57] J. Ray, A. Pinar, and C. Seshadhri. A stopping criterion for markov chains when generating independent random graphs. *IMA Journal of Complex Networks*, 2014.

[58] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

[59] T. J. Santner, B. J. Williams, and W. I. Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

[60] J. Seo, S. Mishra, X. Li, and M. T. Thai. Catastrophic cascading failures in power networks. *Theoretical Computer Science*, 607:306–319, 2015.

[61] T. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2), 2002.

[62] R. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

[63] S. Swarup. A tutorial on synthetic populations, 2017.

[64] X. Wang, S. Wu, K. Wang, X. Deng, L. Liu, and Q. Cai. A spatial calibration model for nanotube film quality prediction. *IEEE Transactions on Automation Science and Engineering*, 13(2):903–917, 2016.

[65] S. Weibel. The Dublin Core: A Simple Content Description Model for Electronic Resources. *Bulletin of the American Society for Information Science and Technology*, 24(1):9–11, November 1997.

[66] L. M. Wein, Y. Choi, and S. Denuit. Analyzing evacuation versus shelter-in-place strategies after a terrorist nuclear detonation. *Risk Analysis*, 30(9):1315–1327, 2010.

[67] B. J. Williams, T. J. Santner, and W. I. Notz. Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, pages 1133–1152, 2000.

[68] X. Yuan, Y. Hu, H. E. Stanley, and S. Havlin. Eradicating catastrophic collapse in interdependent networks via reinforced nodes. *Proceedings of the National Academy of Sciences*, page 201621369, 2017.

[69] A. Zhang and X. Deng. A mixed variance component model for quantifying the elasticity modulus of

nanomaterials. *Technical Report, Virginia Tech*, 2015a.

[70] A. Zhang and X. Deng. A regularized approach to sparse linear discrimination analysis for two-class classification. *Technical Report, Virginia Tech*, 2015b.

[71] H. Zhang, H. Zhang, and M. T. Thai. Chapter 7 a survey on multilayer networks and the applications. In *Big Data in Complex and Social Networks*, pages 181–212. CRC Press, 2016.

[72] Q. Zhang, X. Deng, P. Z. Qian, and X. Wang. Spatial modeling for refining and predicting surface potential mapping with enhanced resolution. *Nanoscale*, 5(3):921–926, 2013.

[73] K. Zuev and M. Beer. Reliability of critical infrastructure networks: Challenges. *arXiv preprint arXiv:1701.00594*, 2017.

Data Management Plan

Data Types. Power networks, communication networks and their coupling structure will be generated as part of the project deliverables. Additionally, representative data and models for human interaction with infrastructure will be generated. The data representation will be designed so as to be extensible to incorporate levels of domain-specific details.

Use of existing standards. The Dublin Core Metadata Element Set [65] will be used to describe common metadata terms across all data formats generated. Domain specific metadata content can be added through the use of appropriate data class extension mechanisms.

Software Sharing, security, IP, re-distribution and derivative data. The data sets developed under this proposal and that are designated for sharing will be made available through a server hosting a digital library, tentatively CINET [1].

The data in CINET is protected and available in a read-only manner which guards against tampering and unauthorized modification. For data sets produced using commercial data or otherwise restricted data, the team will ensure that all permissions and limitations for use of those data sets are adhered to and fully respected.

The license of use, re-use, re-distribution and production of derivative data from the shared data sets is Creative Commons Attribution 4.0 International (CC BY 4.0) [28]. If this license is insufficient for particular uses, an alternate commercial license can be negotiated on a case by case basis.

Preservation of access to data. The released data sets will be available through CINET for two years following the completion of the proposed work. It is anticipated that this will be sufficient time for members of the community to access the data. Moreover, it is expected that, after two years, the research community and/or the proposal team will have generated updated versions of the networks products delivered.