

Identification of Critical Infrastructure via PageRank

Bill Kay*, Hao Lu*, Pravallika Devineni*, Anika Tabassum[†], Supriya Chintavali* and Sangkeun (Matt) Lee*
Oak Ridge National Laboratory* Virginia Tech[†]

Email: {kaybw, luh1, devinenip, chintavalis, lees4}@ornl.gov, anikat1@vt.edu

Abstract—Assessing critical infrastructure vulnerabilities is paramount to arranging efficient plans for their protection. Critical infrastructures are cyber-physical systems that can be represented as a network consisting of nodes and edges and highly interdependent in nature. Given the interdependent nature of critical infrastructures, failure in one node may cause failure in many others resulting in a cascade of failures. In this paper, we propose a node criticality metric that uses Google’s PageRank algorithm to identify nodes that are likely to fail (are vulnerable), nodes whose failure may cascade to many other sites in the network (are important), and nodes that are both vulnerable and important (are critical). We then present a series of experiments to understand how protecting certain critical nodes can help mitigate massive cascading failures. Simulating failures in a real-world network with and without critical node protections demonstrates the importance of identifying critical nodes in an infrastructure network.

I. INTRODUCTION

The Oxford dictionary defines infrastructure as “the basic physical and organizational structures and facilities (e.g. buildings, roads, power supplies) needed for the operation of a society or enterprise” [1], and the executive branch of the United States Government defined *critical* infrastructure as “infrastructures so vital that their incapacitation or destruction would have a debilitating impact on defense or economic security” [2]. While these definitions are consistent with our intuitive understanding, neither one is very precise or technically useful. Further, there are no standard tools for rigorously identifying critical components in a large system of interdependent infrastructures. For example, Hurricane Sandy in 2012 caused enormous power utility damage, which led to the non-operation of major fuel pipelines, telecommunication infrastructure, and water and sewage facilities. Then, the transportation system collapsed due to electric and fuel outages, which led to shutting down health care centers, paralyzing New York [3]. Clearly, society cannot function if large portions of the CI are disrupted or destroyed. Given these heavy interdependencies between critical infrastructures, identifying vulnerable and important units of infrastructure is of great utility to national security interests as well as towards continued operation of a functional society. However, the

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

metrics by which infrastructure can be identified as vulnerable or important are ill defined.

The aim of this paper is to provide a framework through which criticality of infrastructure can be empirically quantified via a so-called *criticality score*. We will argue heuristically that the criticality score is a good measure of how critical an infrastructure asset is. We will further outline and perform experiments on real-world critical infrastructure networks to provide evidence for the fidelity of the criticality score. We will show that protecting critical infrastructure assets drastically reduces the propagation of failures throughout interdependent infrastructures. Armed with the knowledge of which infrastructures are critical, one can allocate resources to prevent failure in critical infrastructure to more efficiently dampen the effects of sporadic infrastructure failure.

In Section II-A, we present interdependent critical infrastructure as a network-based model. In Section II-B, we discuss Google’s PageRank algorithm as an intuitive choice of tool for ranking infrastructure assets according to their vulnerability, importance, or criticality. Related works are introduced in Section III. In Section IV, we introduce a *criticality score* based on PageRank, and introduce experiments designed to demonstrate that assets which receive a high criticality score are indeed critical. In section V, we presents results of experiments performed on the URBANNET critical infrastructure network, a real-world network with 1,941,484 nodes and 21,306,686 edges. The results demonstrate that the criticality score derived from PageRank is indeed indicative of critical infrastructure. In Section VI-C, we recap the connections between PageRank and critical infrastructure, propose subsequent refinements to the experiments to make it adaptable for various real world considerations and discuss future directions.

II. BACKGROUND

We begin this section by providing a rigorous mathematical framework to study infrastructure. We first note that infrastructure assets are dependent on one another, and the failure of one asset can cause a possible failure in another. For example, if an electrical substation provides electricity to a hospital, the failure of the electrical substation is likely to cause the hospital to fail. Although these relationships can be approximated, we seldom have perfect information about dependencies, and so we cannot know how the failure of one infrastructure asset will affect other infrastructure assets. For example, the hospital could have a backup generator, and so failure of the power plant does not necessarily cause the hospital to stop

all operations. We use this notion of dependence to build a directed graph which helps us analyze infrastructure.

A. Infrastructure Network

Given a collection of infrastructure assets A which supply resources to other assets, we construct a directed graph called the *infrastructure network of A*

$$G(A) = (A, E)$$

(or simply G if A is clear from context). The nodes (or vertices) of this graph are the assets A , and for two assets a and b , there is an edge $a \rightarrow b$ (or $(a, b) \in E$) if a supplies a resource to b . Thus, if a fails, b is likely to fail. In this graph, we want to identify *nodes of interest*. There are three types of nodes of interest:

- **Vulnerable nodes:** Nodes which are likely to fail if any other node in the graph fails.
- **Important nodes:** Nodes whose failure causes many other important nodes to be likely to fail.
- **Critical nodes:** Nodes which are both vulnerable and important.

From these we determine a node’s status:

- **Status:** How vulnerable, important, and critical the node is.

Loosely speaking, the status of an asset is dependent on how likely it is to fail if a random asset in the network fails, and how many assets are likely to fail if it fails. a *risk assessment* of an asset is how important it is to an individual based on its status. We use the words network and graph interchangeably in this paper.

Takeaway: A critical site depends on many failure prone assets, and many other important assets depend on it.

B. PageRank

Google’s PageRank algorithm takes a directed graph as input, and assigns to each node the likelihood that a user would have of starting at a random node, walking randomly along edges, and arriving at that node. PageRank is the backbone of the wildly successful Google search engine [4], but has found success in a number of commercial or academic settings such as citation network [5], influence propagation and community detection [6], protein interaction network [7] and personalized recommendation systems [8].

On the other hand, we can run PageRank on the “reverse graph” (a directed graph derived by reversing the edges in the original directed graph). While PageRank tells us which nodes are likely to be reached by a random walk through the graph, the reverse PageRank tells us how likely a random walk is to reach many nodes. For example, in the *internet graph* (nodes are web pages and edges are hyperlinks), a web page v with a high PageRank would tell us that if a user started on a random web page and clicked on random links they would likely wind up at v eventually. A web page v with a high PageRank on the reverse graph would say that if a user started on v and clicked on random links they are likely to see many websites.

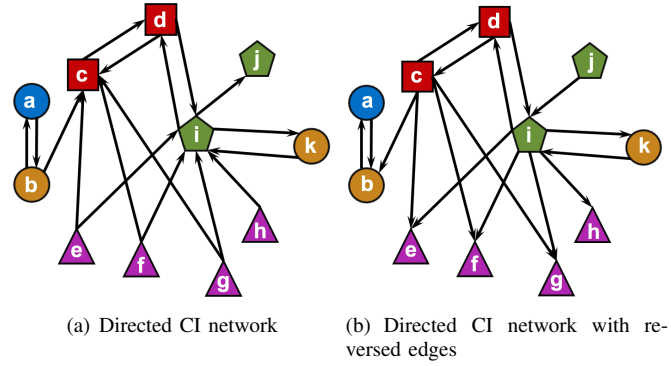


Figure 1. The different colored nodes represent the various critical infrastructure sites and the edges represent the interdependence between these sites. PageRank on the directed graph shows which nodes are influential, while PageRank on the reversed directed graph shows why a node is important.

There are two of nice intuitive properties that can be described as follows:

- Nodes with a high PageRank have many edges pointing towards them from nodes that have many edges pointing towards them and so forth.
- Nodes with a high “reverse” PageRank (i.e., PageRank on the reverse graph) point towards many nodes that point towards many nodes and so forth.

While PageRank determines which nodes are important, it does not specify *why* a node is important. As an illustrative example, Figures 1(a) and 1(b) represent a sample critical infrastructure graph, where the different colored nodes denote the various critical infrastructure sites such as power plants, electric substations, hospitals etc and the edges represent the interdependencies between them. Calculating the PageRank for the network in 1(a) yields d , c and i as the top scoring nodes, making it evident that these are the most influential nodes in the network, correlating with their high connect- edness. Reversing the edges in the above network gives the network 1(b) and applying PageRank to this reversed graph yields b , a and i to be it’s top scoring nodes. Nodes with high reverse PageRank have short paths to many other nodes in the network and frequently the only gateways to these nodes. The highest scoring node b is the only pathway to reaching node a and therefore, a high scoring node in reverse page rank is not necessarily a node having many connections in the network.

Given a user’s needs and an asset’s status, we want to quantify the risk assessment of an asset. We make the following assumptions about the status of an asset:

- An asset is more vulnerable if many of the assets which supply it are also vulnerable, because if any of the assets which supply it fails, it is likely to fail.
- An asset is more important if many of the assets it supplies are important, because if it fails, any of the assets it supplies are likely to fail.

Or, when translated to the language of graph theory:

- **Vulnerable** sites are supplied by many vulnerable sites, and hence many nodes which point towards them have

many nodes pointing towards them and so forth.

- **Important** sites supply many important sites, and hence point towards many nodes which point towards many nodes and so forth.

Combining these notions of vulnerability and importance, our observations about PageRank can be rephrased thusly:

- The nodes which receive a high PageRank score are precisely the nodes we view as vulnerable.
- The nodes which receive a high reverse-PageRank score are precisely the nodes we view as important.
- The average of the PageRank score and reverse-PageRank score is a measure of a node’s criticality (a *criticality score*).

We remark here that while PageRank and reverse PageRank have seen many applications in the literature, it is uncommon to find applications where both metrics are combined into one. Our use-case leans heavily on an equal combination of both the PageRank and the reverse PageRank, and is somewhat unique in this regard. For a survey on PageRank and reverse PageRank, see [9]

C. Graph Construction

1) *DHS HIFLD dataset*: To conduct research on interconnected CI networks, access to real-world data was needed to validate the authors’ assumptions about CI network structure and function. For this purpose, the authors chose the US Department of Homeland Security (DHS) Homeland Infrastructure Foundation-Level Data (HIFLD) dataset [10], which provides open, public domain geospatial data on CIs in a variety of formats. According to their webpage [10], the intent of the DHS HIFLD is “to support research and development efforts focused on community preparedness, resiliency, research, and more.” The HIFLD Open Data Catalog has over 500 layers pertaining to 16 CI sectors.

2) *URBAN-NET critical infrastructure Network in Neo4j*: The authors downloaded the HIFLD data layers [10] as shapefiles, which are a geospatial vector data format. Common geospatial component shapes include points (e.g., power plants and hospitals), multiline strings (e.g., road transportation and electric transmission lines), and polygons (e.g., airports and military bases). The authors used their previously developed software modules to construct a CI network [11] and imported this network into the Neo4j database.

Neo4j is an open-source graph database that is highly scalable and schema-free. It models the data in the form of a graph in which nodes depict the graph entities and edges depict the association of these nodes. Neo4j conveniently provides a declarative query language called Cypher and the Cypher output can be represented as a graph visualization. The current Neo4j database comprises 33 HIFLD layers and 104 interdependency links between these layers. Some layers include electric substations, microwave links, ethanol plants, and hospitals. The database has 1,941,484 nodes and 21,306,686 edges, and the nodes represent infrastructure assets, such as hospitals, wastewater treatment plants, and transmission lines. The current Neo4j instance is housed on Oak Ridge National

Laboratory’s (ORNL’s) Compute and Data Environment for Science (CADES) cloud environment and uses about 2.4 GB of storage space.

In the next section we discuss an experiment which will be used to show that infrastructure assets which receive a high criticality score are indeed critical, and that protecting these strategic assets from failure dampens the cascading effects of sporadic infrastructure failures much better than the protection of random or arbitrary assets. In the subsequent section, the results of this experiment on a real-world dataset are presented.

III. RELATED WORK

In complex networks, due to the cascading failure phenomena resulting from the failure of a small number of nodes, identifying critical nodes is of vital importance to prevent collapse and to maintain the functionality of the network. The methods in this area mainly exploit the network topology and rank the importance of nodes by computing node centrality metrics and node removal [11]. Centrality measures have mainly been used in network-based approaches for vulnerability analysis in power networks [12].

Zhang et al. [13] pointed out that targeted attack causes more damage to interdependent networks than a random attack. Different from a random attack, a targeted attack is simulated by removing nodes with largest scores in terms of topological metrics such as degree, betweenness, and centrality to simulate worst-case disruptions and examine response of the network [14]–[16]. In [17], a collection of network centrality measures are discussed with respect to critical infrastructure. Researchers used modified PageRank algorithm, where they integrated topology and function of a network in infrastructure vertex facility criticality assessment [18]. Differing from cascading failure theory, Shen et al. [19] employed modified weighted PageRank to determine the importance of nodes in both natural gas and electricity network, deployed to increase the security. In this work, the authors’ use PageRank as the primary centrality measure to identify vulnerable and important infrastructure components so they can be protected in case of a failure.

IV. METHODS

In this section, we detail an experiment to test the efficacy of PageRank and reverse PageRank detecting critical infrastructures of A . The results of the experiment will be reported in Section V.

Given an asset A and its corresponding infrastructure network G , we compute the PageRank and reverse PageRank of each node of G . We normalize the scores between 0 and 1, and combine them by averaging the two. Given an asset v , call this score the *criticality score* of v . To each edge $e = (u, v)$, assign a *failure probability* $p \in [0, 1]$ whose interpretation is “If u fails, then v fails with probability p .” In this experiment, every edge of G is assigned the same value of p but further experimentations will allow for variation based on known relationships. For $t, k \in [0, 1]$ and for $N > 1$, we

induce failure in a random t proportion of the nodes of G . We conduct the following experiment:

- 1) Rank the nodes by their criticality scores.
- 2) In two separate instances of the graph, protect (by setting p to 0 on incoming edges) a k proportion of the nodes in the top third (high criticality) and the bottom third (low criticality) of criticality scores. In another instance, protect a *random* k proportion of the vertices.
- 3) Induce a failure in a random t proportion of the nodes of G , and let them propagate for N steps.

A head to head comparison of the total number of nodes which have failed after N steps in each instance will demonstrate the effects of protecting low or high criticality assets vs. protecting a random collection of assets. The more assets that fail, the less effective the protections. We remark here that while we defined criticality score in the context of infrastructure, the notion of a critical node extends naturally to any directed graph which represents dependencies in the same way, and the above experiment can thus be performed in a variety of contexts.

V. RESULTS

We conducted the experiment described in Section IV on the HIFLD Critical infrastructure network. In this experiment, we induce failure in a small number of nodes. We protected 2% of the top third of critical nodes, the top 2% of the bottom third of critical nodes, and a random 2% of nodes according to our PageRank informed criticality scores. We used the following parameters:

- Edge failure probability $p \in \{.1, .5, .8, 1\}$.
- $N = 10$ iterations.
- Initial failure rate $t \in \{.0001, .005, .001, .01\}$.
- Protected proportion $k = .02$.

In the graphs in Figures 2- 5, we have one row of graphs for each of the (fixed) edge failure probabilities. In each row, there is one graph for each initial failure proportion t . The x -axis of each plot is the number of steps the failures have propagated, and the y -axis is the number of vertices in the graph that has failed. Hence, lower values in the plot signify better protected CI networks. The red curves illustrate the protection we get by our high (PageRank based) criticality score, the green curves the low criticality score, and the blue curves represent a random subset of the vertices. We remark that our criticality score out-performs protecting random vertices by a large margin, indicating that criticality is indeed a measure of vertex centrality and a good indicator of which nodes are critical in the intuitive sense. Further, the low criticality nodes perform on par or *worse* than the random nodes. Hence, this criticality score also identifies nodes that are less critical (and does not apply to just the high criticality vertices, which were our main object of interest).

VI. CONCLUSION AND FUTURE DIRECTIONS

A. Conclusion

In this paper, we introduced the notion of critical infrastructure from a heuristic point of view. We then established

a mathematical (network based) framework that allowed us to provide a rigorous definition of vulnerability, importance, and criticality. These definitions were based on the PageRank algorithm, because the interpretation of a high PageRank score translates readily to an intuitive notion of vulnerability while the interpretation of a high reverse-PageRank score translates readily to an intuitive notion of importance. Thus, the average of these two scores translates readily to an intuitive notion of criticality (a mix of vulnerability and importance). We then designed an experiment and conducted it on a real world network that shows that the protection of assets we deem critical does more to ensure the networks' resilience to sporadic infrastructure failure than protecting the same number of nodes at random, or with lower criticality scores. Hence, we conclude that in this large real world network, the PageRank-driven notion of criticality is consistent with our intuitive notion of criticality- the collection of infrastructure assets whose protection does the most to strengthen the entire network.

B. Refinements

The experiment we describe and perform in Sections IV and V showed that in one particular network, the vertices we defined as critical according to the PageRank and reverse PageRank metric are consistent with our intuitive definition of critical. We remark here that there were a number of simplifying assumptions that went into the construction of the network as well as the execution of the experiment. There are two assumptions we would like to highlight as stringent, and possibly not representative of the real-world scenario being modeled. Two such assumptions are:

- 1) Every pair of dependent infrastructures have the same level of dependence.
- 2) The more things an infrastructure depends on, the more likely it is to fail.

Each of these assumptions are a consequence of taking a uniform choice of the failure probability p .

Assumption (1) does not translate directly to all real-world settings, as the degree to which infrastructures depend on each other can vary. For example, we may say that a school depends on the power substation which services it, the water company which provides the potable water, and the news station which broadcasts in its area. While a failure of the first two infrastructures would likely result in a closure of the school, if the news station fails the school can likely still maintain standard operations. The computation of the PageRank algorithm uses the failure probabilities as edge weights, giving more weight to higher failure probabilities (and are thus more likely to cause failure) and less weight to lower probabilities (and are thus less likely to cause failure). In our example, we could give the edges from the power substation to the school and the water station to the school a probability of .9, while we could give the edge from the news station to the school a probability of .1. The power station and water plant would then contribute more strongly to the school's PageRank score. There are a number of ways

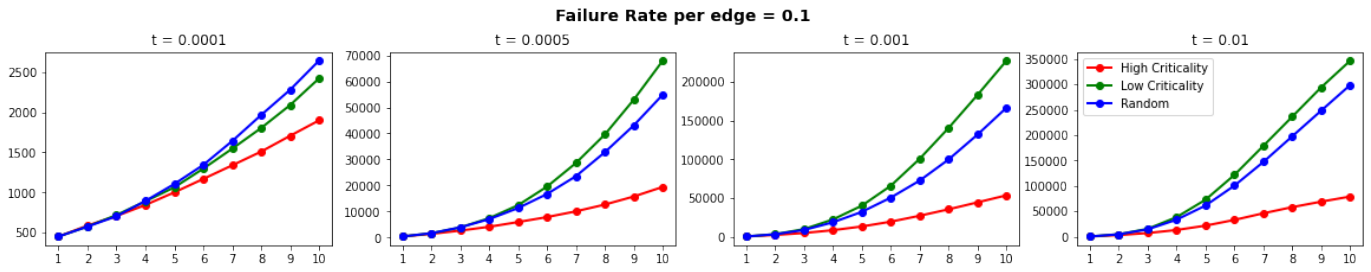


Figure 2.

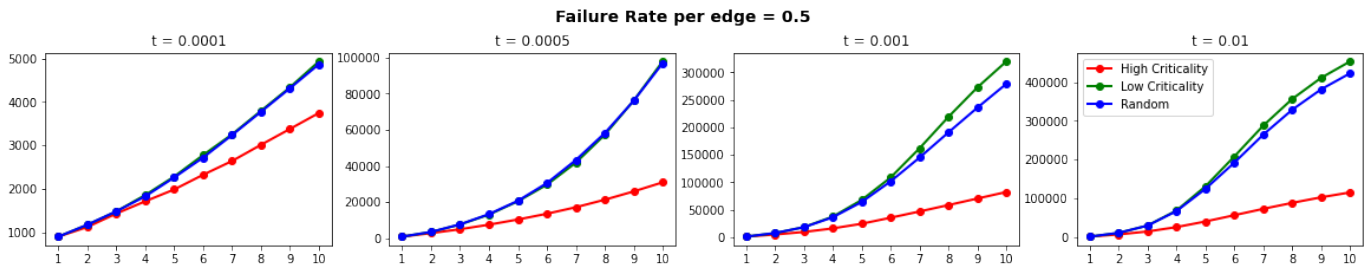


Figure 3.

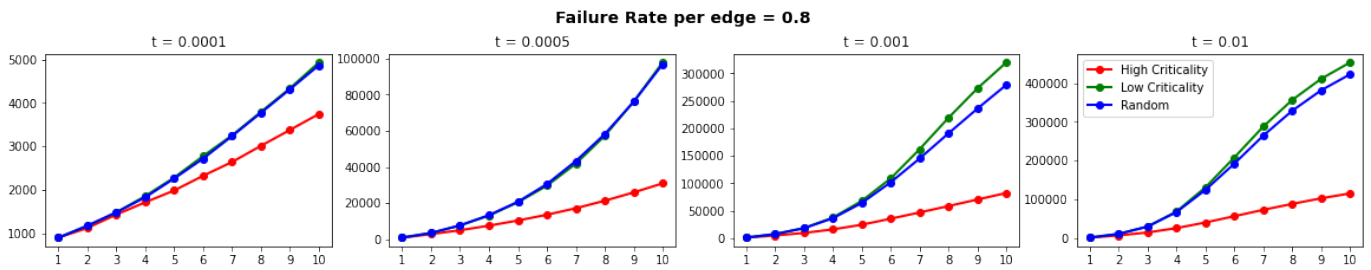


Figure 4.

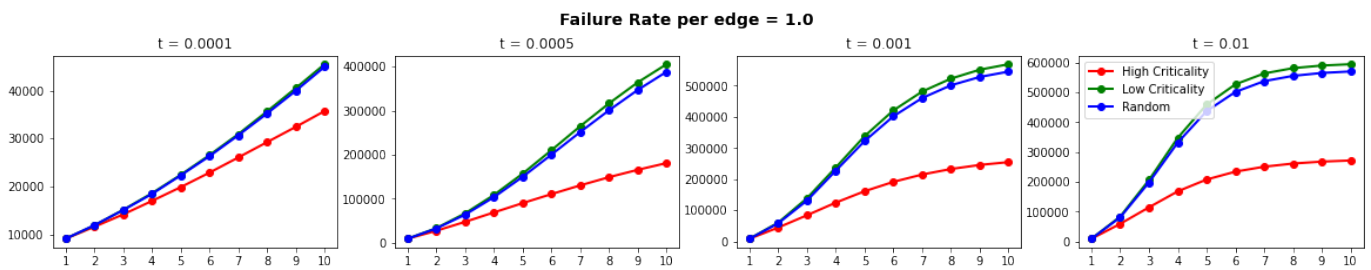


Figure 5.

one can assign the failure probabilities while constructing the model. For example, one could gather historical data on the consequences of different types of infrastructure failure. One could also consider each “type” of source node as having the same level of importance. For example, most infrastructure assets are likely more dependent on a power substation or water station than it is on their local news station. So the infrastructure network could assign failure probability .9 to every edge leaving a power substation or a water station and

.1 to every edge leaving a news station.

Assumption (2) does not translate directly to all real world settings, as sometimes adding dependencies can be viewed as adding redundancies. For example, if a hospital received power from one power substation, and a secondary power substation began providing service as a failsafe, the described experiment would view each of these plants as an equally likely source of failure and the PageRank score would go up (marking the hospital as more vulnerable after the installation of a secondary source of power). As in the previous case, the best possible

scenario would be to gather historical evidence and label the edge failures accurately, but that data is not always readily available. In this case, one could assume that redundancy reduces each of the failure probabilities by some reasonable rule. For example, if n power plants supply a school, then each edge has failure probability $1/n$. In this case, if a school were supplied by a single power plant then the failure probability is 1, two power plants then each failure probability is $1/2$ and so on. PageRank would see the school as having the same weight coming from power plants regardless of the number, but the more power plants the lower the probability that a single power plant can cause failure of the school. Investigation into which method for computing lower edge weights is required.

C. Future Directions

These are the principal directions for future work. First, one can apply any of the refinements of Assumptions (1) or (2) above. Similar analysis would be required, but the models at hand would be more reflective of reality and less prone to faulty analysis via oversimplification. Second, one could compare the criticality score to more sophisticated criticality measures (such as vertices of high degree). Protecting these infrastructures would likely stem the propagation of errors more readily than protecting a random subset of infrastructures. Third, this criticality score could be applied to any directed graph where the edges are constructed to illustrate dependencies. The experiment in Section IV can be applied directly, or tuned to more accurately reflect each usecase.

ACKNOWLEDGEMENT

This material is based upon work supported by DOE's Office of Cybersecurity, Energy, Security, and Emergency Response. This research used resources of CADES at ORNL, which is supported by the US Department of Energy's (DOE's) Office of Science under contract no. DE-AC05-00OR22725.

REFERENCES

- [1] "infrastructure, n." [Online]. Available: <https://www.lexico.com/en/definition/infrastructure>
- [2] B. Obama, "Executive order – improving critical infrastructure cybersecurity," 2013.
- [3] M. Haraguchi and S. Kim, "Critical infrastructure interdependence in new york city during hurricane sandy," *International Journal of Disaster Resilience in the Built Environment*, 2016.
- [4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.
- [5] L. Waltman and E. Yan, "Pagerank-related methods for analyzing citation networks," in *Measuring scholarly impact*. Springer, 2014, pp. 83–100.
- [6] R. Wang, W. Zhang, H. Deng, N. Wang, Q. Miao, and X. Zhao, "Discover community leader in social network with pagerank," in *International Conference in Swarm Intelligence*. Springer, 2013, pp. 154–162.
- [7] X.-Q. Li, T. Xing, and D. Du, "Identification of top-ranked proteins within a directional protein interaction network using the pagerank algorithm: applications in humans and plants," *Curr. Issues Mol. Biol.*, vol. 20, no. 1, 2016.
- [8] F. Jiang and Z. Wang, "Pagerank-based collaborative filtering recommendation," in *International Conference on Information Computing and Applications*. Springer, 2010, pp. 597–604.
- [9] D. F. Gleich, "Pagerank beyond the web," *siam REVIEW*, vol. 57, no. 3, pp. 321–363, 2015.

- [10] D. of Homeland Security, "Dhs homeland infrastructure foundation-level data (hifld)," <https://hifld-geoplatform.opendata.arcgis.com/>, online; accessed 17 September 2020.
- [11] S. Lee, L. Chen, S. Duan, S. Chinthavali, M. Shankar, and B. A. Prakash, "Urban-net: A network-based infrastructure monitoring and analysis system for emergency management and public safety," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 2600–2609.
- [12] G. Oliva, A. E. Amideo, S. Starita, R. Setola, and M. P. Scaparra, "Aggregating centrality rankings: A novel approach to detect critical infrastructure vulnerabilities," in *International Conference on Critical Information Infrastructures Security*. Springer, 2019, pp. 57–68.
- [13] P. Zhang, B. Cheng, Z. Zhao, D. Li, G. Lu, Y. Wang, and J. Xiao, "The robustness of interdependent transportation networks under targeted attack," *EPL (Europhysics Letters)*, vol. 103, no. 6, p. 68005, 2013.
- [14] J. Johansson and H. Hassel, "An approach for modelling interdependent infrastructures in the context of vulnerability analysis," *Reliability Engineering & System Safety*, vol. 95, no. 12, pp. 1335–1344, 2010.
- [15] W. Zhang, Y. Xia, B. Ouyang, and L. Jiang, "Effect of network size on robustness of interconnected networks under targeted attack," *Physica A: Statistical Mechanics and its Applications*, vol. 435, pp. 80–88, 2015.
- [16] L. Duenas-Osorio and S. M. Vemuru, "Cascading failures in complex infrastructure systems," *Structural safety*, vol. 31, no. 2, pp. 157–167, 2009.
- [17] P. Devineni, B. Kay, H. Lu, S. Chintavali, and S. Lee, "Toward quantifying vulnerabilities in critical infrastructure systems," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020.
- [18] C. Zhao, N. Li, and D. Fang, "Criticality assessment of urban interdependent lifeline systems using a biased pagerank algorithm and a multilayer weighted directed network model," *International Journal of Critical Infrastructure Protection*, vol. 22, pp. 100–112, 2018.
- [19] Y. Shen, C. Gu, and P. Zhao, "Structural vulnerability assessment of multi-energy system using a pagerank algorithm," *Energy Procedia*, vol. 158, pp. 6466–6471, 2019.