Pacific Northwest
NATIONAL LABORATORY

PNNL-34896

Inverse Reinforcement Learning based Bayesian Goal Inference Method for Early Nuclear Proliferation Detection

September 2023

Dennis G. Thomas
Zachary J. Weems
Richard E. Overstreet
Benjamin A. Wilson

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34896

Inverse Reinforcement Learning based Bayesian Goal Inference Method for Early Nuclear Proliferation Detection

September 2023

Dennis G. Thomas
Zachary J. Weems
Richard E. Overstreet
Benjamin A. Wilson

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Traditional methods for detection of nuclear proliferation indicators are usually applied after nuclear proliferation has already occurred. There is a need to advance these methods to perform early detection of nuclear proliferation indicators. In this project, we formulated an early detection problem as a sequential, decision-making, goal inference problem based on research publications of authors, to determine whether it is possible to infer whether an author will publish on a research activity before it has occurred. To develop and test our approach, we selected a civil nuclear activity for our case study. We constructed a state-action-state transition graph from publications of authors associated with the activity and the co-authors of their publications, using titles, abstracts, and author publication sequences. We then used inverse reinforcement learning to model the goal-directed behavior of authors in trajectories that terminate at selected goal states. Using a Bayesian formulation, we computed the probability that authors would reach each selected state from partially observed trajectories of their state transitions in their research topic space. The state with the highest probability was selected as the most probable goal state. Based on our results, we found that 60% of the times we can infer the correct goal state early; sometimes the inference is either delayed, or multiple states could be inferred as goal states. Overall, our results show that it is possible to perform early detection of research activities of authors in a nuclear technology area. Further research is necessary to establish a more accurate understanding of how topic modeling, topic space grid discretization, and the extent of overlap among trajectories of different goal states, affect the goal inference results. The methods developed in this work may be used to enhance data-driven methods for early detection of nuclear proliferation indicators.

Summary

Traditional methods for detection of nuclear proliferation indicators are usually applied after nuclear proliferation has already occurred. There is a need to advance these methods to perform early detection of nuclear proliferation indicators. In this project, we formulated an early detection problem based on research publications of authors to determine whether it is possible to infer whether an author will perform a research activity before it has occurred. To develop and test our approach, we selected the development of a civil nuclear activity, such as the construction of the Open Pool Australian Lightwater (OPAL) reactor, as the goal activity to be inferred.

We formulated the early detection problem as a sequential decision-making, goal inference problem, where we represented the publication sequences of authors associated with the goal research activity as a sequence of state-action-state transitions in their research topic space. We used topic analysis based on non-negative matrix factorization (NMF) algorithm to define the research topic space for the OPAL case study. We collected 29,196 Scopus records out of which 270 records formed the primary set of research articles written by five authors (referred to as “coin” authors) of a flagship publication (coin paper) associated with OPAL. The NMF model was trained using the titles and abstracts of these 270 nuclear research articles to define the topic vector space of the authors. We identified ten topics using NMF and mapped the records on to a 10-dimensional topic grid using their NMF-derived topic weight vectors. Each occupied grid cell was then identified as a state in which an author published. We refer the state of the OPAL flagship publication as the coin state. The remaining 28,918 Scopus records formed the secondary set and included articles written by co-authors of articles in the primary set. These were also mapped on to the topic grid using the topic weights computed from the trained NMF model. We defined the actions as the difference in the number of years it took for each author to move from one state to the other between two consecutive publications. This resulted in 403 author trajectories with their state-action-state transitions represented as a Markov decision process (MDP).

The goal inference problem was to infer whether each of the five coin authors would publish in the coin state after observing a partial trajectory of their state transitions. Using inverse reinforcement learning, we selected four sets of author trajectories associated with four terminal (goal) states (the coin state being one of the goal states) and modeled the goal-directed behavior of authors towards publishing in each goal state. We then developed a Bayesian formulation to compute the probability that each coin author will publish in the coin state given their partial trajectory of state transitions. Our results show that we are able infer the coin state as the correct goal state for 3 of the authors after 2 steps of observation. For two authors, the inference was delayed until the author was two to three steps from reaching the goal state. Specifically, for one of them there were two goal states, including the coin state, that were equally probable goal states. Thus, for some trajectories, multiple goal states may be possible, and the actual goal state may not be realized until the terminal step of the trajectory.

This work presents the first attempt at using nuclear research articles for early detection of research activities of authors in a nuclear technology area. We showed results for one case study associated with a civil nuclear research activity. Further research is necessary to establish a more accurate understanding of how topic modeling, topic space grid discretization, and the extent of overlap among trajectories of different goal states, affect the goal inference results. The methods developed in this work may be used to enhance data-driven methods for early detection of nuclear proliferation indicators.

Acknowledgments

This research was supported by the **Mathematics for Artificial Reasoning in Science Initiative**, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

1.0 Introduction

Most methods used for identifying nuclear proliferation indicators such as chemical signatures and nuclear activities, often enable detection of nuclear proliferation after an entity has acquired special nuclear materials or a specific nuclear technology (Sheffield 2020). Nuclear research articles have been used to identify early proliferation indicators such as influential research entities and technical expertise levels of a country in a nuclear technology area (Kas et al. 2012, Chatterjee et al. 2023). With recent advancements in data science, computing, and artificial intelligence, it may be possible to perform early detection of nuclear activities in a technology area from scientific and technical documents (Sheffield 2020, Alexander et al. 2020). Detection of nuclear proliferation indicators from data is limited due to partial observability, sparse and unlabeled information, and confounding signals from multiple concurrent activities. In this work, we consider the problem of early detection of research activities in a technology area from nuclear research articles to demonstrate a proof-of-concept approach for early detection of nuclear proliferation.

We cast the early detection problem as a sequential decision-making, goal inference problem, where the objective is to predict the probability that an agent (e.g., a country) will pursue a nuclear activity (e.g., building a reactor) from partially observable sequences of activities, using inverse reinforcement learning (IRL) and Bayesian goal inference methods. The problem is motivated from the problem of driver destination and route prediction from partial trip trajectories; where, given a list of driver trip trajectories on a discretized geographical map, the goal is to learn the behavior of these drivers and then to predict the destination of similar behaving drivers given their partial trip trajectories (Krumm and Horvitz 2006, Ziebart et al. 2008). In our work, we formulate the goal inference problem for nuclear research activities of author state-action-state transitions, created based on their temporal sequence of their publications. We develop a computational approach using inverse RL to model the authors' behavior towards publishing in a particular goal state (location) in the topic vector space of publications in a nuclear research area. We then develop a Bayesian formulation to predict whether an author will publish in a particular state associated with a nuclear technology or research activity.

In the basic RL framework, an agent learns to follow a sequence of state-action-state transitions to reach a particular goal state depending on the transition probability and the rewards it accumulates along the way. The sequence of transitions is modeled as a Markov decision process (MDP). The behavior of the agent to reach a goal state is modeled using a reward function that depends on the state, state-action, or state-action-state features in the MDP environment. In the inverse RL framework, the actual goal state and the associated rewards are not known, but only an observed partial sequence of state transitions of an agent. The objective of the IRL-based Bayesian goal inference problem is to infer the most probable goal state of the agent, given its partial sequence of state transitions. An IRL algorithm is used to solve the reward function that models the goal state behavior of the agent using historical data of known trajectories that terminate at the same goal state. Based on the rewards, a policy – the probability of performing an action when the agent is in a state – is computed and used in a Bayesian formulation to predict the most probable goal state.

2.0 Approach

To demonstrate the application of our IRL-based Bayesian goal inference approach, we formulated a goal inference problem surrounding a well-documented civil nuclear activity – the construction of the Open Pool Australian Lightwater (OPAL) reactor in Australia. The OPAL reactor is a 20-MW multipurpose reactor, used for producing radioisotopes for cancer detection and treatment, and neutron beams for fundamental materials research (ansto.gov). It went critical in August 2006 and was officially opened in 2007. Our approach to the OPAL activity goal inference problem is to use a flagship publication associated with the OPAL reactor as the goal activity to be inferred from the publication sequences of authors doing research in the nuclear reactor domain. We identified the flagship publication by searching Scopus for records with keywords "Opal" and "reactor", found in the title and abstract. Since the OPAL reactor went critical in 2006, search results from 2004 through 2008 were considered. The chosen paper was titled "Novel cryogenic engineering solutions for the new Australian research reactor OPAL" (Olsen et al. 2008). We refer to this flagship paper as the "coin" paper and the authors of this paper as coin authors. The paper was written by nine authors, out of which five authors had Scopus ID's associated with a previous publication history.

There are 3 main steps in our IRL-based Bayesian goal inference approach, as illustrated in Figure 1:

1. Construct state-action-state transition graph from publications using title, abstract, and author publication sequences.

2. Compute rewards for a set of author state-action-state trajectories that terminate at the same state, using an IRL algorithm.

3. Calculate goal probability for a partially observed trajectory of state transitions using a Bayesian formulation.

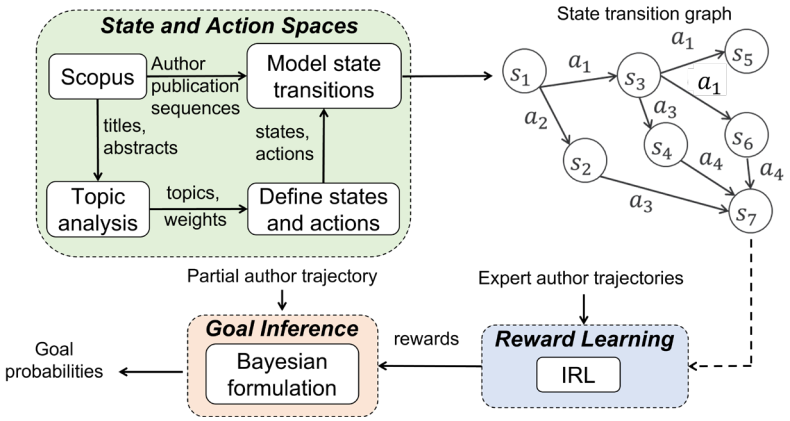

Figure 1 IRL-based Bayesian goal inference approach for author publication trajectories.

2.1 Construction of state-action-state transition graph

After identifying the coin paper, we constructed a state-action-state transition graph to represent the authors' MDP environment for IRL. This involves defining the state space, action space, and state-action-state transition probabilities. There could be multiple ways of defining states and

actions. We defined the states as the cells of a *K*-dimensional grid that represented the topic vector space of all the papers published by the five coin authors through the end of 2008. Based on the Scopus search records, there were 278 coin-authored papers including the coin paper. From each Scopus record, we extracted the titles, abstracts, publication dates, and author ID's of each paper. To define the topic vector space of the coin authors, we applied Nonnegative Matrix Factorization (NMF) on the titles and abstracts of 270 coin-authored papers – 8 papers that were written in 2008 (before or after the coin) were omitted from the NMF training set to minimize overlap of topics with the coin paper around its time of publication. We used the NMF algorithm implemented in the scikit-learn package (Pedregosa et al. 2011) for the topic analysis.

We determined an optimal number of ten topics (*K* = 10) to represent the topic vector space. The optimal *K* value of 10 was identified using the Kneedle algorithm (Satopaa et al 2011), after considering various NMF hyperparameter settings and regularization methods. Thus, the resulting 270 topic weight vectors, each of length 10, were mapped on to a 10-dimensional rectangular grid. The grid cell of each paper is considered the *state* of the paper or of its authors. To introduce noise in the OPAL MDP environment, we applied the trained NMF model to calculate the 10-dimensional topic weight vectors for 28,918 Scopus records of the papers written by the co-authors of all the 278 papers. Thus, a total of 29,196 papers were mapped onto a 10-dimensional grid.

States in the author MDP environment: The number of grid cells (states) in the OPAL MDP environment will depend on the choice of the grid cell spacing along each grid dimension and the number of occupied grid cells. We divided each grid dimension into 5 intervals, with outer intervals having width equal to half the width of the inner intervals. Mapping the records by their topic weights on the grid resulted in 603 grid cells occupied with at least one record. Thus, the state space of the OPAL consisted of 603 states.

Actions in the author MDP environment: We defined the actions as the difference in the number of years it took for authors to move from one state to another. For example, if an author published in t years from one state to another state, then a directed edge is drawn for the action t from the first to the second state. Our action definition resulted in 15 actions (the maximum year difference observed was 15 and the minimum was 0) in state-action-state transition graph. If the year difference was 0, we used the month and day information from the publication dates to determine the edge direction along each transition. We computed the state-action-state transition probabilities based on the number of authors who moved along each edge. All self-loops due to an author publishing in the same state in consecutive steps were ignored during construction of the state-action-state transition graph.

2.2 Reward learning using IRL

All state-action-state transitions ($s - a - s'$) were modeled as first order MDPs with transition probabilities $T(s, a, s')$, state rewards $R_1(s)$, and state-action rewards $R_2(s, a)$. We used the exact maximum entropy (ExactMaxEnt) IRL algorithm developed by Snoswell et al. (2020) to compute the rewards for a group of authors (experts) whose trajectories terminate at the same goal state. Since the state and state-action features are not known, we fitted the rewards for all states and state-action pairs by using the ExactMaxEnt algorithm to predict and match the average state and state-action visitation frequencies observed in the expert trajectory set (training set). Starting with a uniform distribution of random values for the rewards, we performed 3000 iterations to update the rewards until convergence. Convergence was based on the reward gradients and correlation between the observed and predicted state and state-action visitation frequencies.

2.3 Bayesian formulation for goal inference

The objective of the goal inference is to predict whether an author with a partially observed trajectory of state transitions, will pursue a publication in the coin state. For this purpose, we first model the behavior of authors who have previously visited the coin state, by computing the rewards based on their trajectories using IRL. Unlike driver trip trajectories, author trajectories do not have a definite starting point from which they start to have an intent to publish in a goal state. Hence, we can choose any starting point, which will determine the length of the trajectory to the goal state. We also assume that the intent to publish in the goal state begin at the previous state of each step along the trajectory. For demonstrating the goal inference, we also model the behavior of authors who have trajectories terminating at other selected goal states in the dataset. Thus, the objective of our Bayesian formulation is to determine which of the selected goal states, including the coin state, is an author with a partial trajectory most likely to publish in. The trajectories in each set associated with the goal state can be of different lengths (number of states). In our analysis, we trim the trajectories to a specified length, L, from the starting state. This length is determined by the shortest author trajectory in the coin author trajectories.

Let's consider a partial trajectory of t state transitions, denoted as $(s_{k-1}, s_k)_{k=1}^{t}$. Let there be n possible goal states, denoted as G_i , where $i \in [1,2, \dots, n]$. The probability of an agent publishing in a goal state G_i , based on the goal state reward policy, π_i, given t observations of state transitions of the author's partial trajectory, is formulated as,

$$P_{\pi_i}(G_i|(s_{k-1}, s_k)_{k=1}^{t}) = \frac{\sum_{k=1}^{t} P_{\pi_i}(G_i|(s_{k-1}, s_k))}{\sum_{j=1}^{n}\sum_{k=1}^{t} P_{\pi_j}\left(G_j|(s_{k-1}, s_k)\right)}. \quad (1)$$

The term $P_{\pi_i}(G_i|(s_{k-1}, s_k))$ is the posterior probability of reaching state G_i in $L-1$ steps with reward policy π_i , given the transition (s_{k-1}, s_k):

$$P_{\pi_i}(G_i|(s_{k-1}, s_k)) = \frac{P_{\pi_i}((s_{k-1}, s_k)|G_i)P_{\pi_i}(G_i|s_{k-1})}{\sum_{j=1}^{N} P_{\pi_i}\left((s_{k-1}, s_k)|S_j\right) P_{\pi_i}(S_j|s_{k-1})}. \quad (2)$$

Here, N is the total number of states in the topic grid MDP, $P_{\pi_i}(G_i|s_{k-1})$ is the prior goal probability based on all observed paths of length $\leq L$ from s_{k-1} to G_i in the expert trajectory set, and $P_{\pi_i}((s_{k-1}, s_k)|G_i)$ is the likelihood given by the formula,

$$P_{\pi_i}((s_{k-1}, s_k)|G_i) = \frac{P_{\pi_i}(s_k \to G_i)}{P_{\pi_i}(s_{k-1} \to G_i)}. \quad (3)$$

$P_{\pi_i}(s_{k-1} \to G_i)$ and $P_{\pi_i}(s_k \to G_i)$ are the total path probabilities based on all possible paths (of steps 1 to $L-1$) from s_{k-1} and s_k to the state G_i, respectively.

3.0 Results

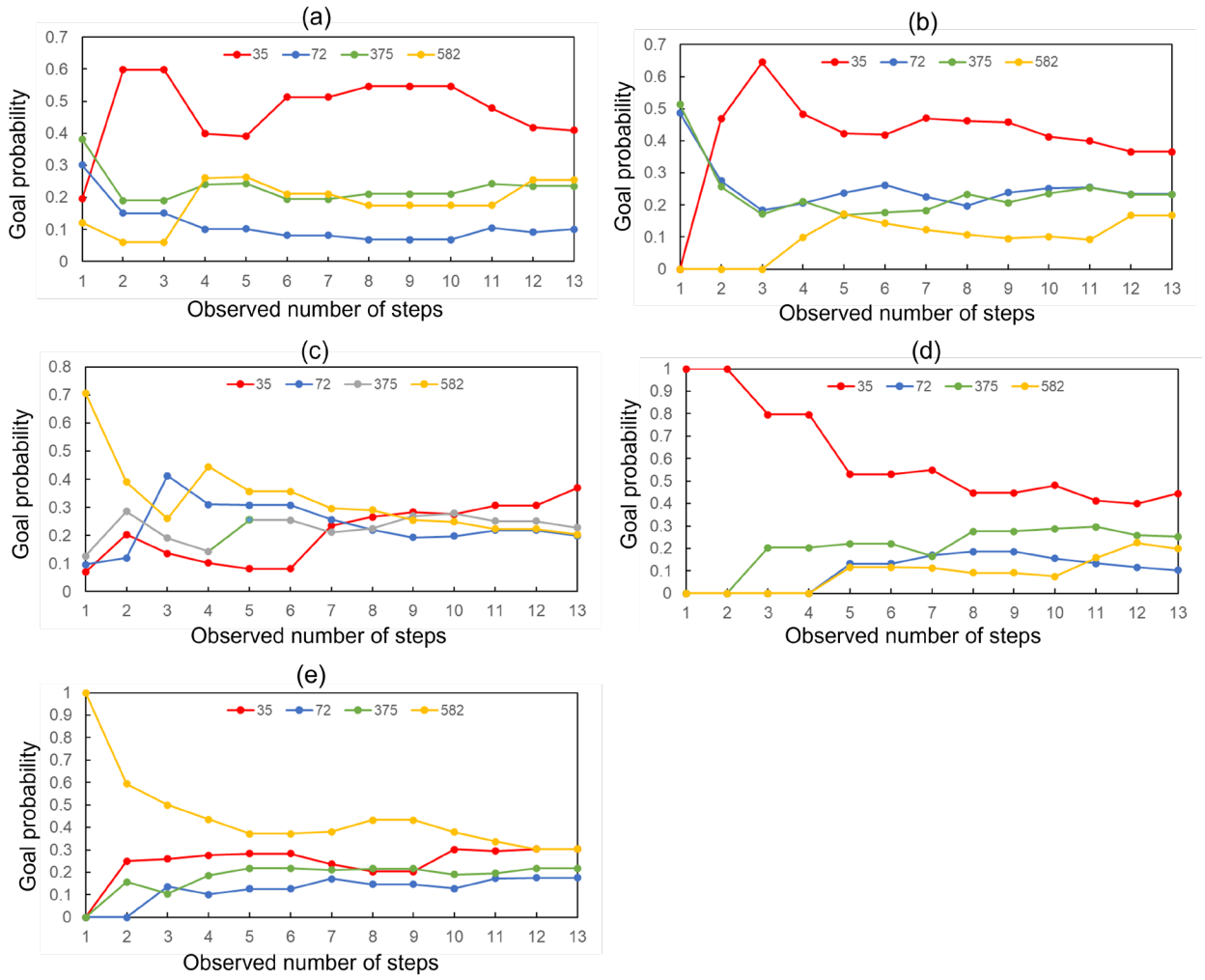

Figure 2 Goal probabilities for states 35, 72, 375, and 582, using the respective reward policy as more number of steps are observed along each coin author trajectory (a-e).

There were 403 author trajectories in the dataset, which spanned across 603 states in the topic author MDP. To test the performance of the IRL-based Bayesian goal inference formulation, we considered four goal states, identified here by numbers as 35, 72, 375, and 582. State 35 is the coin state and the other goal states were arbitrarily selected such that their highly weighted topics were different from each other and that of the coin state. To learn the reward-based policy for modeling the goal state-directed behavior of the authors towards each goal state, we created a training set of author trajectories that terminate at each goal state. The number of trajectories in the training set for goal states 35, 72, 375, and 582 were 16, 14, 18, and 15, respectively. The five coin author trajectories (in no particular order) were excluded from the training set to use them as a test set for external validation. The number of steps each coin author took from their starting state in the trajectory to the coin state were 69, 14, 31, 42, and 78, respectively. Since the lowest number of steps was 14, we ignored states that were away from the goal state by more than 14 steps, so that all the trajectories in the training sets and in the 5 coin trajectory set had 14 steps between the starting state and their respective goal state.

Thus, all the trajectories used for IRL were of length $L = 15$. The IRL simulations were run for each trajectory set and the results converged in 3000 iterations. The root mean squared values of the reward gradients were below 0.004 and the linear correlation coefficients between the observed and predicted values of both state as well as state-action visitation frequencies for each trajectory set were close to 1. After learning the reward policy for each goal state, we performed goal inference test on the five coin author trajectory set.

For each coin author trajectory we computed the goal probability values for all four goal states, using Equation (1), as more number of steps are observed along the trajectory. The state with the highest probability value was selected as the most probable goal state. Figure 2 shows the goal probability values for the four goal states. The results indicate that for three out of the five trajectories (Figures 2(a), 2(b), and 2(d)), the probabilities are the highest for the coin state after the first or second step of observation. Thus, the coin state is inferred as the correct goal state at least 11 steps before the author reaches the coin state. For the third coin author trajectory the inference of the coin state is delayed by 11 steps (Figure 2(c)). It appears that state 582 is the goal state until 8 steps. In the 9^{th} and 10^{th} step, either states, 35 or 375, can be the goal state. But after 10 steps, it appears the coin state is the most probable state. In the case of the fifth trajectory, one would infer state 582 as the most probable goal state (Figure 2(e)). In step 13, although the probability was slightly higher by 0.0008 for the coin state, both the coin state and state 582 are equally probable. Thus, for some trajectories, multiple goal states may be possible, and the actual goal state may not be realized until the terminal step of the trajectory. Thus the coin state was the inferred goal state in 60% of the trajectories (i.e., 3 out of 5) after 2 steps of observation. After 11 steps of observation, the percentage rose to 80% (4 out of 5).

4.0 Conclusion

In this project, we developed an IRL-based Bayesian goal inference approach to determine the goal state of an author from partial trajectories of state transitions in their research topic space. We showed results for one case study for early detection of a civil nuclear research activity using nuclear research articles. Further research is necessary to establish a more accurate understanding of how topic modeling, topic space grid discretization, and the extent of overlap among trajectories of different goal states, affect the goal inference results. This work presents the first attempt at using nuclear research articles for early detection of research activities of authors in a nuclear technology area. The methods developed in this work may be used to enhance data-driven methods for early detection of nuclear proliferation indicators.

5.0 References

Sheffield, A. "Developing the Next Generation of AI Systems to Push the Detection of Foreign Nuclear Proliferation further 'Left of Boom'." *Countering Weapons of Mass Destruction Journal* (2020).

Alexander, Francis J., Tammie Borders, Angie Sheffield, and Marc Wonders. Workshop Report for Next-Gen AI for Proliferation Detection: Accelerating the Development and Use of Explainability Methods to Design AI Systems Suitable for Nonproliferation Mission Applications. No. BNL-221083-2021-FORE. Brookhaven National Lab.(BNL), Upton, NY (United States); Idaho National Lab.(INL), Idaho Falls, ID (United States); National Nuclear Security Administration (NNSA), Washington, DC (United States), 2020.

Kas, Miray, Alla G. Khadka, William Frankenstein, Ahmed Y. Abdulla, Frank Kunkel, L. Richard Carley, and Kathleen M. Carley. "Analyzing scientific networks for nuclear capabilities assessment." *Journal of the American Society for Information Science and Technology* 63, no. 7 (2012): 1294-1312.

Chatterjee, Samrat, Dennis Thomas, Daniel Fortin, Karl Pazdernik, Benjamin Wilson, and Lisa Newburn. "Dynamic Network Analysis of Nuclear Science Literature for Research Influence Assessment." *ESARDA Bulletin* 65, no. PNNL-SA-166748 (2023).

Krumm, John, and Eric Horvitz. "Predestination: Inferring destinations from partial trajectories." In *International Conference on Ubiquitous Computing*, pp. 243-260. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

Ziebart, Brian D., Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. "Maximum entropy inverse reinforcement learning." In *Aaai*, vol. 8, pp. 1433-1438. 2008.

Olsen, S. R., S. J. Kennedy, S. Kim, J. C. Schulz, R. Thiering, E. P. Gilbert, W. Lu, M. James, and R. A. Robinson. "Novel cryogenic engineering solutions for the new Australian Research Reactor OPAL." *In AIP Conference Proceedings*, vol. 985, no. 1, pp. 299-306. American Institute of Physics, 2008.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

Satopaa, Ville, Jeannie Albrecht, David Irwin, and Barath Raghavan. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior." *In 2011 31st international conference on distributed computing systems workshops*, pp. 166-171. IEEE, 2011.

Snoswell, Aaron J., Surya PN Singh, and Nan Ye. "Revisiting maximum entropy inverse reinforcement learning: new perspectives and algorithms." *In 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 241-249. IEEE, 2020.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov