Pacific Northwest
NATIONAL LABORATORY

PNNL-32797

Synthetic Data and Graph Generation for Modeling Adversarial Activity

Final Project Report

February 2022

Sumit Purohit
Patrick S Mackey
Joseph A Cottam
Madelyn Dunning
George Chin

U.S. DEPARTMENT OF ENERGY

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

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

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;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-32797

Synthetic Data and Graph Generation for Modeling Adversarial Activity

Final Project Report

February 2022

Sumit Purohit
Patrick S Mackey
Joseph A Cottam
Madelyn Dunning
George Chin

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

The Data and Graph Generation for Modeling Adversary Activity project developed a methodology along with scalable graph modeling and generation tools to produce realistic large-scale background activity graphs with embedded adversarial activity pathways. Pacific Northwest National Laboratory performed as a data generator and evaluator to produce datasets for use by activity tasks TA2 and TA3 performers. The released datasets included background activities, activities of interest, and merged graphs that establish ground truth, and filtered versions of merged graphs that have been separated into individual data types. Templates representing activity patterns for TA3 performers to detect were also released as part of each deliverable.

Acronyms and Abbreviations

AIDA	Active Interpretation of Disparate Alternatives
GDF	Graph Definition File
INDRA	Integrated Network and Dynamical Reasoning Assembler
ITeM	Independent Temporal Motif
KG	knowledge graph
MAA	Modeling Adversary Activity
MSB	Money Service Business
NLP	natural language processing
NYC	New York City
PNNL	Pacific Northwest National Laboratory
QLiG	Query Like a Graph
RDF	Resource Description Framework
SME	subject matter expert
SPG	Semantic Property Graph
UIUC	University of Illinois at Urbana-Champaign
WMD	weapon of mass destruction

Contents

Figures

1.0 Introduction

The Data and Graph Generation for Modeling Adversary Activity (MAA) project developed a methodology along with scalable graph modeling and generation tools to produce realistic large-scale background activity graphs with embedded adversarial activity pathways. Pacific Northwest National Laboratory (PNNL) performed as a data generator and evaluator to produce datasets for use by MAA tasks TA2 and TA3 performers. The released datasets included background activities, activities of interest, and merged graphs that established ground truth, and filtered versions of merged graphs that have been separated into individual data types. Templates representing activity patterns for TA3 performers to detect were also released as part of each deliverable.

Critical data production tasks for the project included development of weapons of mass destruction (WMDs) and other transitional domain scenarios or use cases, preparation and provision of real-world background and adversary activity data, and research and development of scalable graph integration and modeling algorithms for the construction of large-scale background activity graphs and the embedding of adversarial activity patterns into background activity graphs.

2.0 Graph Generation and Data Release Schedule

MAA Phase	Dataset Name (Major Versions)	Release Date	Short Description	URL
P1	PNNL-V1	08/15/2017	Six multichannel (C1, C2, and C3) graphs: three background graphs without signal graph patterns and three corresponding background graphs. They were stochastically generated such that a graph with a signal is not simply an insertion of the signal graph pattern into the background graph, but a different graph altogether.	https://maa-graphs.labworks.org/Versions/V1/V1.aspx
P1	PNNL-V2	10/31/2017	Multichannel (C1, C2, and C3) graphs with three sizes: 7K (~7k nodes per channel), 50K (~50k nodes per channel), and 100K (~100k nodes per channel).	https://maa-graphs.labworks.org/Versions/V2/V2.aspx
P1	PNNL-V3	3/15/2018	Multichannel (C1, C2, C3, C4, and C5) graph with 120 unique signal graphs and 12 stochastically generated background+signal.	https://maa-graphs.labworks.org/Versions/V3/V3.aspx
P1	PNNL-V4	6/15/2018	Multichannel (C1, C2, C3, C4, and C5) graph with 12 small data sets of approximately 10K nodes per channel with one signal copy embedded in each dataset, 12 large datasets of approximately 1M nodes per channel with 10 signals embedded in each dataset, and two noise model post-processing codes that can operate on the provided datasets to degrade their fidelity. There are four node types in each dataset and six edge types.	https://maa-graphs.labworks.org/Versions/V4/V4.aspx
P1	PNNL-V5	9/15/2018	Similar to V4 with five node types in each dataset and seven edge types.	https://maa-graphs.labworks.org/Versions/V5/V5.aspx
P1	PNNL-V6	1/15/2019	Update of V5.	https://maa-graphs.labworks.org/Versions/V6/V6.aspx
P1	PNNL-V7	5/15/2019	Multichannel graphs with five medium datasets on the order of 100K people nodes with five signals embedded in each data set, and one large dataset on the order of 10M people nodes with 25 signals embedded in it.	https://maa-graphs.labworks.org/Versions/V7/V7.aspx

MAA Phase	Dataset Name (Major Versions)	Release Date	Short Description	URL
P1	PNNL_Real_World_Alignment	3/1/2019	A real-world multilayer network dataset constructed from data collected by network science researchers Matteo Magnani and Luca Rossi. The data represent friendship or following relationships from three social media services.	https://maa-graphs.labworks.org/Versions/RW/Real-world.aspx
P1	PNNL-Milan	7/15/2019	A multisource dataset of urban life in the city of Milan and the Province of Trentino (G. Barlacchi et al., Scientific Data 2, Article number: 150055 (2015) https://www.nature.com/articles/sdata201555). The Milan dataset include two layers: • Phone (represented by eType 0) • SMS text messaging (represented by eType 7)	https://maa-graphs.labworks.org/Versions/Milan/Milan.aspx
P2	PNNL_AIDA_V1	10/15/2019	A set of exemplar Resource Description Framework (RDF) knowledge graphs (KGs) generated as part of DARPA-AIDA M18 evaluation phase describing Ukrainian political crisis.	https://maa-graphs.labworks.org/Versions/AIDA/V1.aspx
P2	PNNL_AIDA_V2	1/9/2020	Semantic property graphs (SPGs) are generated using DARPA-AIDA M18 evaluation phase describing Ukrainian political crisis.	https://maa-graphs.labworks.org/Versions/AIDA/V2.aspx
P2	PNNL_NYC_V3	7/31/2020	SPGs that combines graph generated using the artificial intelligence for data analytics (AIDA) pipeline and additional transactional sources such as: Venmo, Reddit, Publications.	https://maa-graphs.labworks.org/Versions/AIDA/V3.1.0.aspx
P2	PNNL_MSB_V1	11/30/2020	KGs and templates for a single channel of financial transactions.	https://maa-graphs.labworks.org/Versions/MSB/V1.0.0.aspx
P2	PNNL_MSB_V2	4/7/2021	Geographic Data File (GDF) version of data that was previously released as csv. The GDF contains de-identified data from the financial transactions. All personally identifiable information attributes have been de-identified using cryptographic hashes. Only exact matching is possible. When relevant, some attributes have been de-	https://maa-graphs.labworks.org/Versions/MSB/V2.1.0.aspx

MAA Phase	Dataset Name (Major Versions)	Release Date	Short Description	URL
			identified using bloom filters that allow for some comparison of de-identified values for similarity.	
P2	PNNL_Covid_V3.0	1/6/2021	Graph data and template for the COVID-19 science-based use case. The KG is generated using the Harvard Integrated Network and Dynamical Reasoning Assembler (INDRA) and the University of Illinois at Urbana-Champaign (UIUC) Blender Lab datasets. Both of the raw datasets are integrated to form a large KG in GDF format	https://maa-graphs.labworks.org/Versions/Covid/Covid-v3.aspx
P2	PNNL_Covid_V3.3	8/9/2021	Updated COVID biological KG using INDRA and PyBel. Structural MAA query schema QLiG to construct new templates.	https://maa-graphs.labworks.org/Versions/Covid/Covid-v3.3.0.aspx
P2	PNNL_NYC_V4	2/26/2021	Dataset that expands on previous MAA NYC V3 data releases. Includes a new data channel (sensor measurements) that comes from the DARPA SIGMA+ program. Also has additional contextual information from Reddit, Venmo, and news articles.	https://maa-graphs.labworks.org/Versions/NYC/V4.0.0.aspx

3.0 Graph Generation Approaches

PNNL developed a diverse set of graphs for MAA performers to develop network alignment (TA2) and subgraph matching (TA3) algorithms. This was achieved using different generation methodologies, primarily generating two different classes of the graph. Phase 1 of the program focused on the mathematical formulation of the alignment and subgraph matching problems. This led to the generation of large synthetic structure-only graphs that were used to evaluate TA2 and TA3 algorithms without using personally identifiable information for privacy-preserving requirements.

3.1 Phase 1 MAA Graph Generation

PNNL developed new generative models to simulate multichannel, heterogeneous activity graphs that provide early indicators of acquire, fabricate, proliferate, and/or deploy weapons of mass terror. Since the state of the art supports single-channel, homogeneous models, PNNL developed a network simulation framework to model social networks, communication, procurements, and co-authorship. Modeling multichannel networks simultaneously with correlated channel attributes at scale compounds complexity. Including adversarial signals across channels in a large and high-fidelity synthetic graph compounds that difficulty further.

We developed a framework for the synthetic background graph and template embedding process. The background graph models large-scale benign activities observed for the entire population of the graph. In contrast, the template represents the activities of interest performed by a small set of individuals or groups. PNNL made the following significant contributions to the research field of synthetic graph generation:

1. New generative model to simultaneously produce a correlated multichannel graph.
2. High-fidelity template embedding in the background graph because of inherently simultaneous generation of background and signal activities as a graph.
3. Scalable generative model using fault-tolerant, distributed algorithms and execution environment.

As shown in Figure 1, PNNL developed a graph-generation pipeline to produce multiple datasets in Phase 1. For the background graphs, we used activity and organizational patterns to construct the background, with some decisions strongly influenced by subject matter expert (SME) input on the needs of adversarial activity support. At the end of Phase 1, we generated a large multichannel graph with communications (Phone and Email), procurement, co-authorship, travel, financial transactions, and demographic data channels. We used different hyperparameters to configure the pipeline such as the distribution of motifs in the communication graph, purchasing behaviors, and the travel probability, temporal rhythms, spatial distribution, etc.

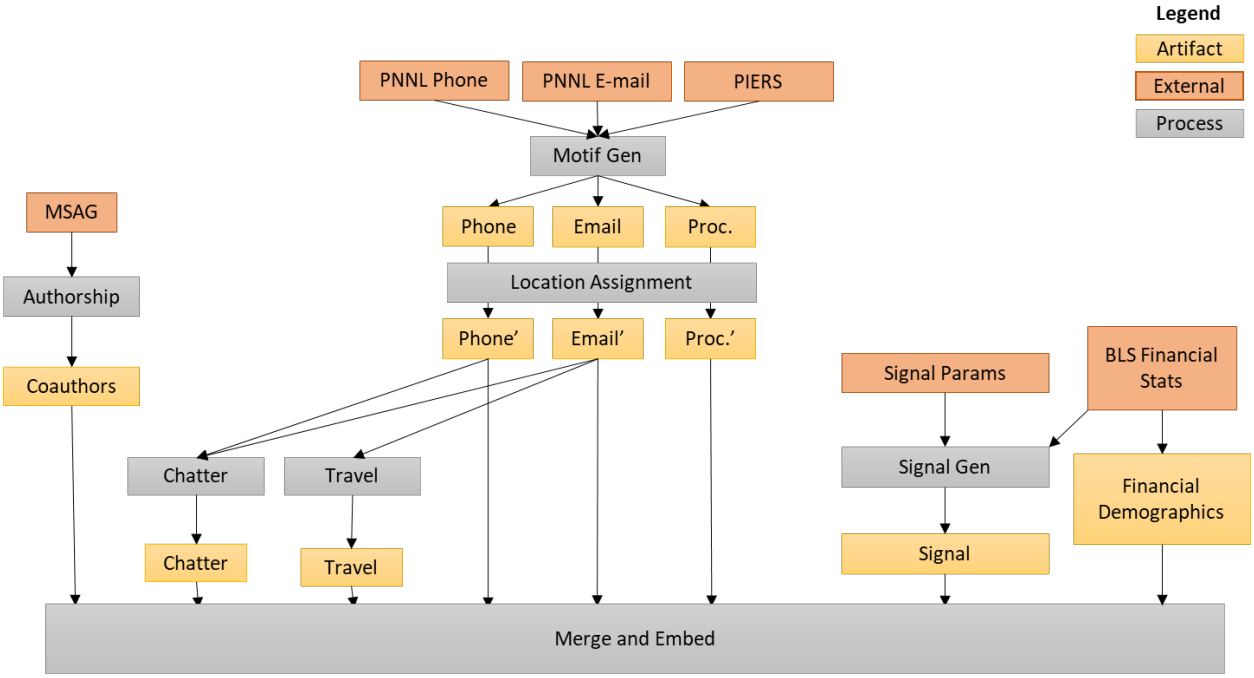

Figure 1. Graph Generation Pipeline

For the multichannel communication network, PNNL developed a new generative model that used Independent Temporal Motif (ITeM) as shown in Figure 2. ITeMs are small temporal structures that can model local transactions observed in a system and be used to represent the temporal evolution of the system. We used real-world aligned data for email and phone communication with PNNL's Living-Lab efforts and computed temporal evolution patterns such as motif formation duration, edge arrival delay, etc. of the ITeM motifs to simulate the network. We also introduced rhythmic temporal patterns to simulate circadian rhythm. The memory-less nature of the generative model scales up to billions of edges. We generated phone and email networks simultaneously using this model.

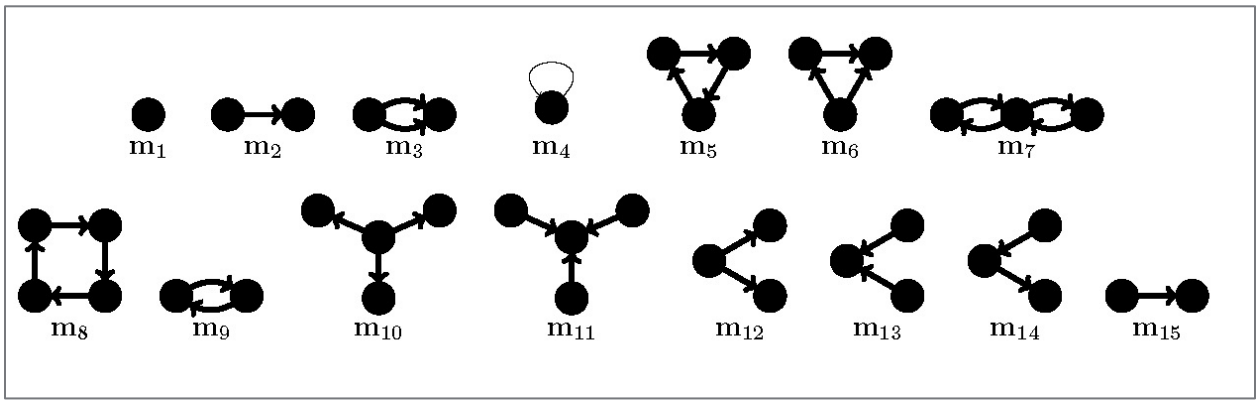

Figure 2. ITeM Library of Motifs

We also developed a chatter model that can represent a regional increase in communication catalyzed by actions such as a terrorist attack, an election, or the death of somebody notable. Such events often manifest themselves as regional changes to communication behavior, typically an increase. The chatter model replicates a localized increase in a way that respects network

topology. Multiple versions of resultant multichannel graphs with embedded signals were released to MAA performers as shown in Figure 3.

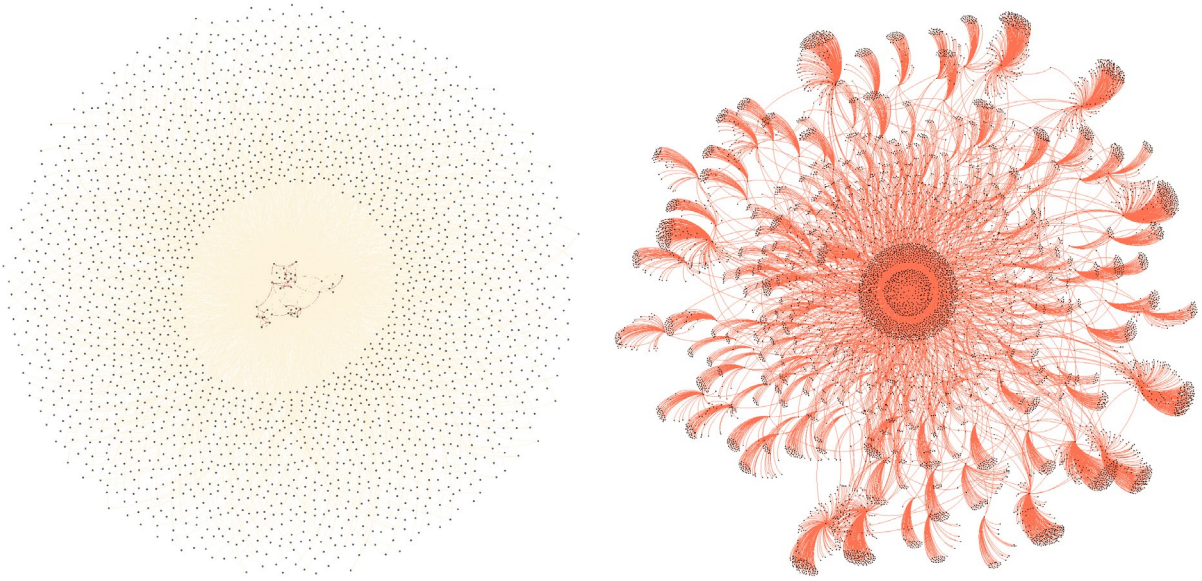

Figure 3. PNNL Graphs for MAA Phase 1

3.2 Phase 2 MAA Graph Generation

MAA Phase 2 focused on semantic, attributed KGs, in contrast to structure-only multichannel graphs in Phase 1. PNNL used real-world data sources to produce different KG datasets for TA2 and TA3 performers. The WMD KG continued to be the primary dataset that involved formal evaluation. In addition, PNNL generated different transition use-case KGs and challenge problems for money-laundering and biological pathway discovery domains.

The Phase 2 KG generation process can be categorized into three different approaches:

1. Transformation from existing KGs
2. Transactional KG construction from heterogeneous data sources
3. KG generation based on domain-specific ontologies and sources.

PNNL started with KG generation based on DARPA AIDA M18 datasets. AIDA M18 data are about the Ukrainian-Russian political conflict of 2013–2014. PNNL released the PNNL_AIDA_V1 dataset that included a set of exemplar RDF KGs generated from the AIDA M18 dataset. RDF graph model and corresponding serialization were found to be non-optimal for MAA problems of network alignment and subgraph matching. PNNL developed SPGs to represent AIDA data into the labeled property graph format. SPG is a logical projection of reified RDF into the property graph model. Instead of RDF serialization such as Turtle, PNNL started using GDF format, which is a tabular representation of nodes and edges in the graph. Based on PNNL recommendations to use SPG, MAA Phase 2 used non-RDF graphs for TA2 and TA3 challenge problems. PNNL developed a cloud-scale framework for SPG generation of the AIDA-based graph as shown in Figure 4.

PNNL continued to generate WMD activity graphs as its primary deliverables. Traditionally, KGs are used to describe metadata about entities and provide additional context to a target application. Many real-world domains also involve temporal interactions between entities in addition to the metadata. Modeling these attributed transactions is a critical requirement when using KGs in complex real-world applications, such as modeling adversarial activities. PNNL developed new capabilities to generate transactional KGs that are heterogeneous activity graphs representing a large background graph and multiple instances of small graphs signifying activities of interest for WMD indicators. We define transaction as an interaction between entities with additional metadata about the interaction.

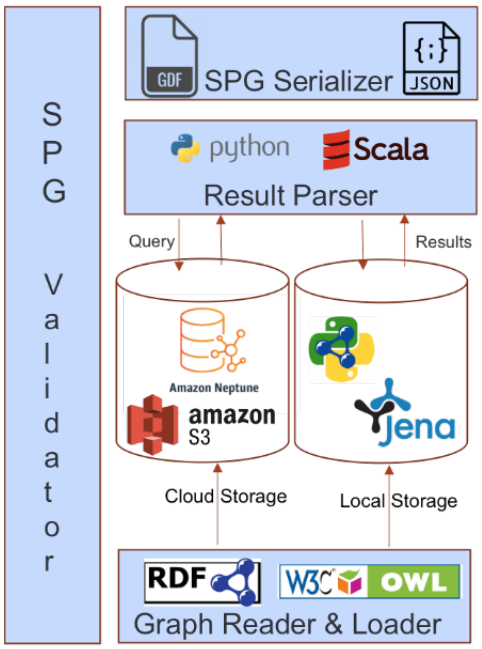

Figure 4. SPG Generation Framework

PNNL_NYC_V3 and PNNL_NYC_V4 were the WMD dataset generated using the transactional KG generation capabilities PNNL developed for Phase 2. We put together several publicly available data sources that cover events that occurred within New York City (NYC) from January 2018 to December 2019. We looked at a wider range of data sources that provide valuable related information closely associated with WMDs, including internet forum discussions related to extremism and weapon building (as well as benign topics associated with neighborhoods in NYC), and bibliographic data associated with topics related to various kinds of WMDs. For WMD KGs, we used the following data sources with increasing size and improved quality from V3 to V4:

- News articles
- Reddit
- Venmo
- Bibliographic/publication data.

Figure 5 shows the framework PNNL proposed to develop the transactional KG graph from heterogeneous sources. We leveraged a natural language processing (NLP)-based entity and relation-extraction capabilities developed as part of the AIDA program and customized them for MAA-WMD scenarios using a cloud-scale NiFi pipeline. We also used WikiData as the reference knowledge base to ground entities and events in the NYC graphs. To generate NYC and MSB datasets, PNNL also developed baseline capabilities to align entities and events in transactional KGs. These alignment capabilities are based on name rarity and frequency observed in the source dataset. We plan to improve it going forward while working on transition problems.

The final approach of Phase 2 graph generation dealt with KG generation based on domain-specific ontologies and sources. For the MAA COVID19 use case, PNNL generated a KG describing biological pathways from sources such as Harvard INDRA and UIUC Blender Lab datasets. They use NLP pipelines to extract pathway knowledge from the COVID-19 Open Research Dataset that contains over 60,000 scientific publications about COVID-19. We constructed a large KG using the causal assertions related to SARS-CoV-2 infections using node types such as Gene, Chemical, Protein, Disease, Biological Process, Reaction, Complex, and

Abundance. The edge types in the KG represent the functional relationship between nodes. PNNL focused on addressing data modeling challenges to integrate new taxonomies and data sources.

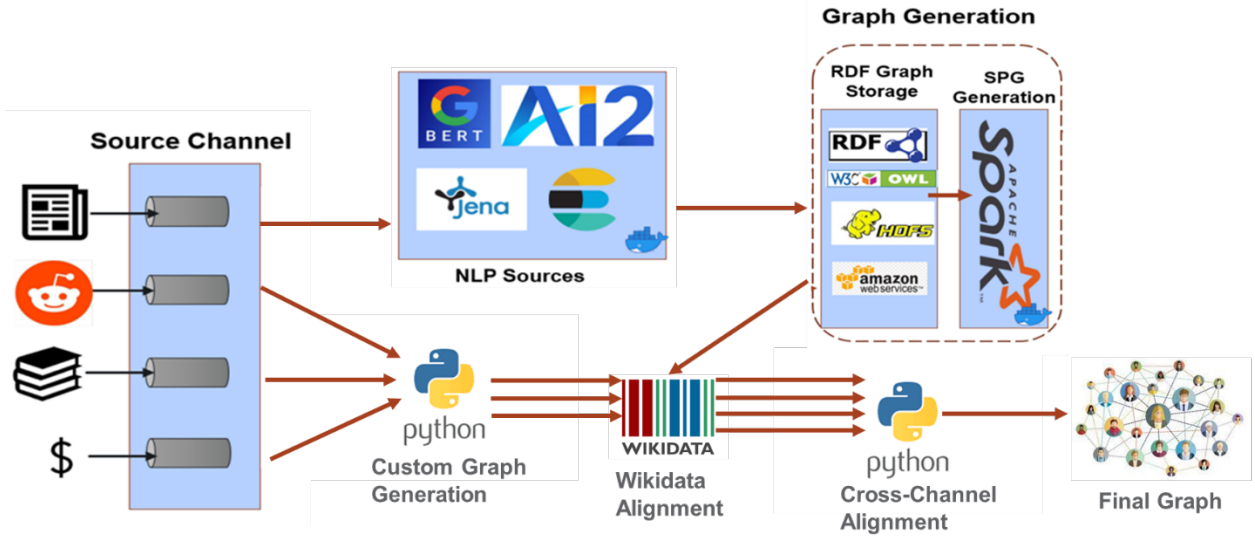

Figure 5. Notional Transactional KG Generation Framework

3.3 MAA Template Generation

In addition to background generation, PNNL developed new approaches to generate and embed signals of interest into the background. For MAA Phase 1, signal embedding was primarily driven by the motif-based generative model and the chatter model we developed. Both approaches ensure the signal is not immediately apparent in the background. To measure the accuracy of signal embedding we developed the signal-to-clutter ratio as shown in Figure 6. This ratio was provided as an intractable integral approximated via Monte Carlo simulation using consistent machine learning as shown in Figure 7.

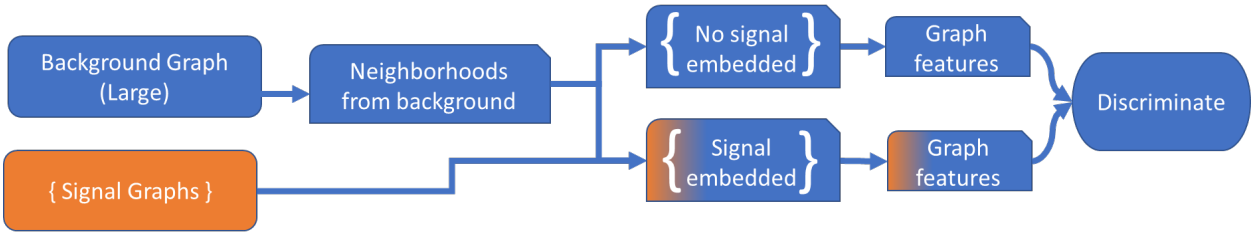

Figure 6. MAA Phase 1 Signal-to-Clutter Measurement

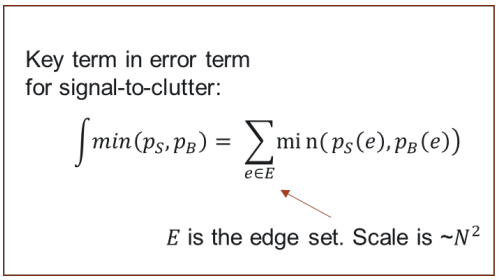

Figure 7. Signal-to-Clutter Formulation

MAA Phase 2 signal generation approaches varied based on the use case involved. For the WMD use case, PNNL continued to focus on realistic signal embeddings and generating core signals using the same pipeline as the background graph with some scenario-specific customizations. We curated a collection of data sources that describe WMD activities and generated the signal using the NLP pipeline. At a high level, we used the following process to generate and embed signals:

1. Scenario description and identification of key entities, types, and relationships
2. Signal generation that describes activities as a graph
3. Templatization (i.e., parameterization) of the signal.

For MSB and COVID-19 use cases, instead of generating signals that are a proxy of real-world events, we worked with SMEs to identify true-positive signals in the background. We further parameterize them to generate signals. PNNL also explored auto-generation of templates for the MAA_AIDA use case where it developed graph neural network capabilities to learn key node types based on a trained set of signals to generate additional templates.

3.4 MAA Query Language

Starting with MAA Phase 1, no specific query language was decided for the subgraph matching problems. PNNL developed graph-based query approaches in both Phases 1 and 2. For the structure-only Phase 1 graphs, PNNL represented query templates as a fine-grained graph where each entity and interaction was listed as a graph node and edges. This approach continued for the initial challenge problems in Phase 2. As the program moved to KGs across different use cases, earlier bottom-up approaches of query specifications were found limited in expressivity and usability. The piecemeal approach also does not support approximate subgraph matching because the query requires every node and edge to be present in the result. SMEs need not describe every possible detail of the activity (bottom-up approach) because of the unavailability of information, fuzzy nature of the activity, and group structure are observed.

Instead, PNNL proposed a top-down *a*pproach to specify a template using high-level node/edge membership, cardinality, and constraints. PNNL developed Query Like a Graph (QLiG), a new query specification that defines graph-based components (e.g., node, edge, path, structure) as the basic building blocks of the query language. QLiG allows SMEs to naturally describe domain-specific patterns using higher level concepts such as group, path, and constraint. PNNL open-sourced the specification where the vocabulary includes the following: node, edge, structure, path, and constraints (attributed, structure, type, membership). QLiG also defines functional categories of edges that can be reused in other parts of the query. This powerful capability allows SMEs to easily find patterns in a KG as it simplifies the MAA queries as shown in Figure 8.

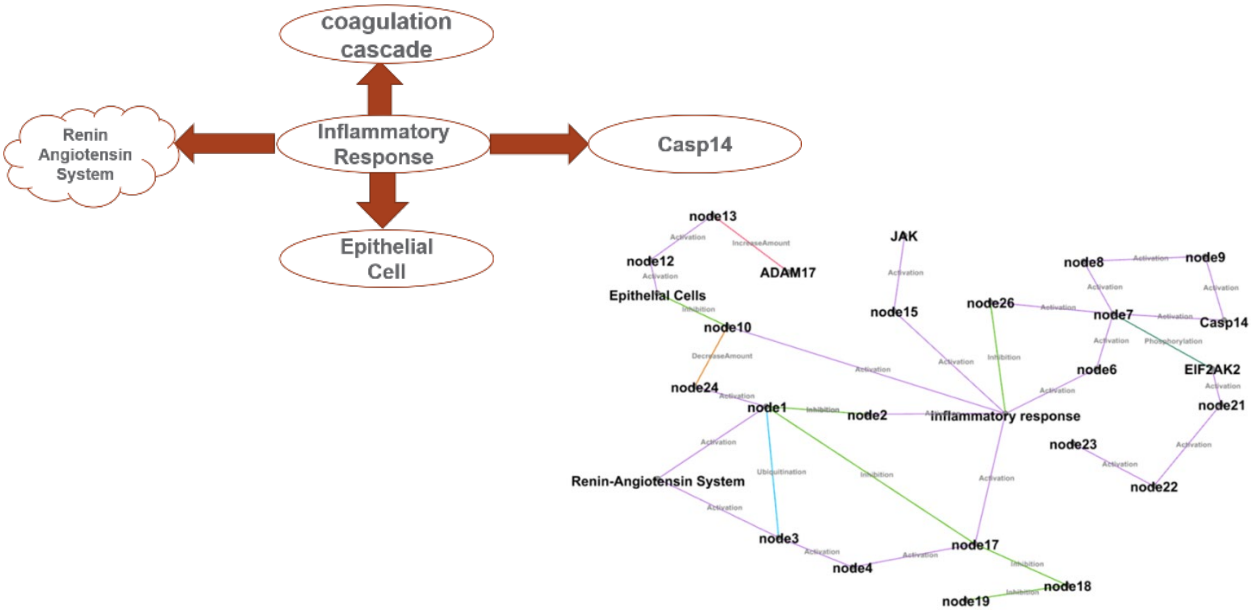

Figure 8. QLiG Representation (right) of Bottom-up Graph Query (left)

4.0 Impact

4.1 MAA Transition Partners

Concepts that were developed for the NYC WMD use case were applied in a pair of pilot studies for two different potential government transition partners. In both cases we interviewed analysts at these agencies to determine what kind of graphs and templates could be used in fuzzy subgraph matching to help solve the particular problem they were faced with. For the first agency , we created graphs of the order of 5M nodes and 11M edges with one to two data channels, and six different templates. We collected and summarized the TA3 performers' results and demonstrated them to the government partner. They were able to confirm that some of the matches were indeed true positives. We followed this with alignment tasks for the TA2 performers that were relevant for the analysts' needs.

For our second transition partner, we collected data from a relevant source and created a set of templates that could be tested against for an operational use case. They were able to confirm that the results were relevant and sent us additional multichannel data to work with. The results of this work have been considered valuable enough that they continue working with us as a transition partner. In addition to template matching, we also provided experimental results in graph alignment and graph generation.

4.2 MAA Use Cases

In addition to the primary MAA use case (NYC WMD) and transition partner use cases, we also generated a biological KG related to SARS-CoV-2 pathways. We curated existing sources such as Harvard INDRA and UIUC Blender Lab datasets. The source data were generated using NLP on a corpus of scientific publications. The primary objective of this use case was to validate MAA graph generation and analytics methodologies including developed capabilities on a diverse application domain. In contrast to NYC and transition partner use cases, MAA COVID-19 KG has no activities and group structures observed in the graph. Biological pathways involve genes, chemicals, proteins, etc. instead of transactions. This led to query templates generation where the templates were a lot different than the rest of the use cases. We worked with computational biologists to construct high-impact scientific questions as a graph template. QLiG query schema was extensively used to describe higher order fuzzy templates. While working with MAA TA3 performers, we identified that the large, fuzzy templates lead to high-combinatorial solution space and must be addressed by specific and narrow constraints on the graph structures. MAA performers were able to identify multiple matching pathways including ground truth. With the COVID-19 use case, we also highlighted the importance of graph analytics to provide richer contextual information and augment the key words-based search approaches such as Google search.

4.3 Publications

PNNL published multiple papers, reports, datasets, and invention disclosures in synthetic graph generations, temporal graph characterization, evaluation of graph algorithms, graph neural network applications, and human computer interfaces.

Cottam JA, NC Heller, CL Ebsch, RD Deshmukh, PS Mackey, and G Chin. 2020. "Evaluation of Alignment: Precision, Recall, Weighting and Limitations." In *IEEE International Conference on Big*

Data (Big Data 2020), December 10-13, 2020, Atlanta, GA, 2513 - 2519. Piscataway, New Jersey:IEEE. PNNL-SA-156949. doi:10.1109/BigData50022.2020.9378064

Cottam JA, S Purohit, PS Mackey, and G Chin. 2018. "Multi-Channel Large Network Simulation Including Adversarial Activity." In IEEE International Conference on Big Data (Big Data 2018), December 10-13, 2018, Seattle, WA, 3947-3950. Piscataway, New Jersey: IEEE. PNNL-SA-138688. doi:10.1109/BigData.2018.8622305

Dunning MP and S Purohit. 2019. "Higher Order Temporal Analysis of Global Terrorism Data." In GTA³ 3.0: The 3rd workshop on Graph Techniques for Adversarial Activity Analytics. PNNL-SA-148159.

Ebsch CL, JA Cottam, and G Chin. 2021. "Evaluating the subgraph matching problem in the presence of categorical constraints." In *The 5th workshop on Graph Techniques for Adversarial Activity Analytics (GTA³ 4.0)*. PNNL-SA-167487.

Ebsch CL, JA Cottam, NC Heller, RD Deshmukh, and G Chin. 2020. "Using Graph Edit Distance for Noisy Subgraph Matching of Semantic Property Graphs." Presented by C.L. Ebsch at the 4th Workshop on Graph Techniques for Adversarial Activity Analytics, Atlanta, Georgia. PNNL-SA-158130.

Joaristi M, S Purohit, RD Deshmukh, and G Chin. 2020. "Data-Driven Template Discovery Using Graph Convolutional Neural Networks." In *IEEE International Conference on Big Data (Big Data 2020), December 10-13, 2020, Atlanta, GA*, 2534-2538. Piscataway, New Jersey: IEEE. PNNL-SA-156967. doi:10.1109/BigData50022.2020.9378318

Mackey P, K Porterfield, E Fitzhenry, S Choudhury, and G Chin. 2018. “A Chronological Edge-Driven Approach to Temporal Subgraph Isomorphism.” In IEEE International Conference on Big Data (Big Data 2018), December 10-13, 2018, Seattle, WA, 3947-3950. Piscataway, New Jersey:IEEE. PNNL-SA-138688. doi:10.1109/BigData.2018.8622305

Mackey PS, WP Smith, MP Dunning, S Purohit, CJ Larimer, MJ Orren, and JA Cottam, NC Heller, CL Ebsch, TM Langlie-Miletich, and G Chin. 2021. "Multilayer Subgraph Matching for Adversarial Activity Detection." Abstract submitted to Graph Fest 2021, Baltimore, Maryland. PNNL-SA-161036.

Orren M, P Mackey, N Heller, and G Chin. 2020. “Multi-channel Entity Alignment via Name Uniqueness Estimation.” In *IEEE International Conference on Big Data (Big Data 2020), December 10-13, 2020, Atlanta, GA*, 2534-2538. Piscataway, New Jersey: IEEE. PNNL-SA-156967. doi:10.1109/BigData50022.2020.9378318

Purohit S, F Shelobolin, L Holder, L Holder, and G Chin. 07/19/2021. "Temporal Analysis of Epidemiology indicators and Air Travel Data for Covid-19." Abstract submitted to SIAM Conference on Applied and Computational Discrete Algorithms, "Online Conference", Washington. PNNL-SA-160489.

Purohit S, L Holder, and G Chin. 2018. "Temporal Graph Generation Based on a Distribution of Temporal Motifs." In 14TH INTERNATIONAL WORKSHOP ON MINING AND LEARNING WITH GRAPHS (MLG 2018), August 20, 2018, London, United Kingdom. PNNL-SA-134797.

Purohit S, NC Heller, G Chin, CL Ebsch, RD Deshmukh, and JA Cottam. 2021. "Difficulty metrics for subgraph isomorphism." Abstract submitted to Complex Networks 2021, Online Conference, Spain. PNNL-SA-166380.

Purohit S, PS Mackey, JD Zucker, A Bohra, RD Deshmukh, and G Chin. 2021. "QLiG: Query Like a Graph For Subgraph Matching." Presented by S. Purohit at Artificial Intelligence & Knowledge Engineering 2021, "Online Conference", United States. PNNL-SA-168736.

Purohit S, PS Mackey, WP Smith, MP Dunning, MJ Orren, TM Langlie-Miletich, RD Deshmukh, A Bohra, TJ Martin, DJ Aimone, and G Chin. 2021. "Transactional Knowledge Graph Generation To Model Adversarial Activities." In *2021 IEEE Big Data Conference*. PNNL-SA-167380.

4.4 Open Source Software Releases

https://github.com/temporal-graphs/STM: ITeM in a temporal network to represent various complex systems.

https://github.com/temporal-graphs/QLiG: QLiG (pronounced cleeg) a graph-based query specification to perform subgraph matching in a Labeled Property Graph.

https://github.com/pnnl/temporal_subgraph_isomorphism: C++ source code for performing temporal subgraph matching against attributed directed temporal networks.

5.0 Recommendations and Lessons Learned

Graph generation has been a challenging research problem in both Phase 1 and Phase 2 with distinct requirements. Graph generation is a DARPA hard problem because of fidelity, scalability, usability, and reproducibility requirements. It has been observed as a combination of graph modeling, statistical analytics, adversarial behavioral modeling, distributed computing, etc. PNNL lists the following recommendations based on its experience in MAA Phases 1 and 2 as graph generators:

1. Structure-only graphs (background and signal) do not represent the real-world fidelity required for an operational system. Structure-only graphs make the subgraph matching problem an NP-Complete problem and lead to the combinatorial explosion of search space.
2. At the same time, synthetic generation for the structure-only graph is a tractable problem because it does not deal with real-world multidimensional parameter/attribute dependencies.
3. Irrespective of the mode of graph generation (structure only vs. attributed), simultaneous generation of the signal is recommended for the accurate embedding of signals into the background graph.
4. Template generation is still primarily an SME-driven process and further research is required to automate the generation of credible signals and templates. This will also lead to faster adoption of the graph analytics system by domain experts.
5. Standardization of query language is key to usability and expressivity of the pattern matching capability. Structural query language development is an active research area. High-level query language leads to complex and realistic templates and expedites the algorithm development for improved performance and generality for real-world applications.
6. State-of-the-art NLP capabilities produce high-quality entity extractions but relationship extraction is still an open research area. MAA deals with activity graphs which also requires extractions of accurate roles and relationships between entities. PNNL leveraged DARPA AIDA capabilities and extended them for the MAA NYC use case, but it can be further improved.
7. Lack of attributes and contextual information limits graph algorithm performance. Availability of rich attributes from the real-world dataset is a challenge and in the absence of additional information, network alignment and subgraph matching suffer large solution space and produce suboptimal solutions.

Pacific Northwest National Laboratory

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

1-888-375-PNNL (7665)

www.pnnl.gov