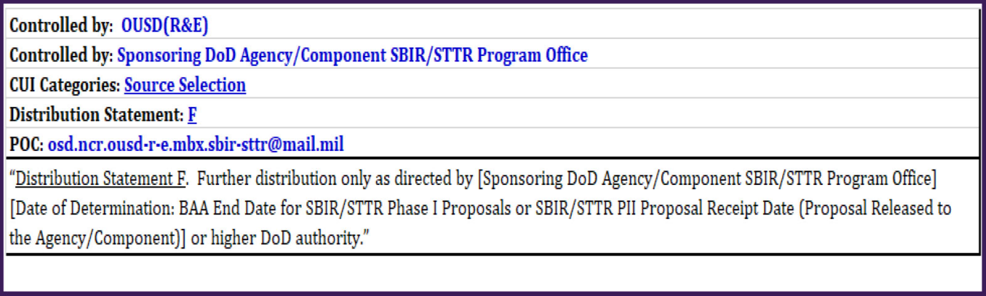
CUI

ATTENTION



Controlled by: OUSD(R&E)	
Controlled by: Sponsoring DoD Agency/Component SBIR/STTR Program Office	
CUI Categories: Source Selection	
Distribution Statement: F	
POC: osd.ncr.ousd-r-e.mbx.sbir-sttr@mail.mil	

"Distribution Statement F. Further distribution only as directed by [Sponsoring DoD Agency/Component SBIR/STTR Program Office] [Date of Determination: BAA End Date for SBIR/STTR Phase I Proposals or SBIR/STTR PII Proposal Receipt Date (Proposal Released to the Agency/Component)] or higher DoD authority."

ATTENTION

All individuals handling this information are required to protect it from unauthorized disclosure.

Handling, storage, reproduction, and disposition of the attached document(s) must be in accordance with 32 CFR Part 2002 and applicable agency policy.

Access to and dissemination of Controlled Unclassified Information shall be allowed as necessary and permissible to any individual(s), organization(s), or grouping(s) of users, provided such access or dissemination is consistent with or in furtherance of a Lawful Government Purpose and in a manner consistent with applicable law, regulations, and Government-wide policies.

Standard Form 901 (11-18)
Prescribed by GSA/ISOO | 32 CFR 2002

CUI

Small Business Innovation Research(SBIR) Program - Proposal Cover Sheet

Disclaimer

Knowingly and willfully making any false, fictitious, or fraudulent statements or representations may be a felony under the Federal Criminal False Statement Act (18 USC Sec 1001), punishable by a fine of up to $10,000, up to five years in prison, or both.

SBIR Phase I Proposal

Proposal Number:	**F244-0001-0097**
Proposal Title:	**S.M.A.R.T. L.I.N.K.S. - "Streaming Modular Adaptation and Real-time Transformation for Learning and Inference in Knowledge Structures"**

Agency Information

Agency Name:	**USAF**
Command:	**AFMC**
Topic Number:	**AF244-0001**

Firm Information

Firm Name:	**thatDot, Inc.**
Address:	**421 SW 6th Ave. Suite 300, Portland, OR 97204-1622**
Website:	**https://www.thatdot.com**
UEI:	**YDL7H2F7BH93**
CAGE:	**8EQU1**
SBA SBC Identification Number:	**002455077**

Firm Certificate

OFFEROR CERTIFIES THAT:

1. It has no more than 500 employees, including the employees of its affiliates.	**YES**
2. Number of employees including all affiliates (average for preceding 12 months)	**12**
3. The business concern meets the ownership and control requirements set forth in 13 C.F.R. Section 121.702.	**YES**
4. Verify that your firm has registered in the SBAS Company Registry at www.sbir.gov by providing the SBC Control ID# and uploading the registration confirmation PDF:	**SBC_002455077**

Supporting Documentation:

- SBC_002455077.pdf

5. It has more than 50% owned by a single Venture Capital Owned Company (VCOC), hedge fund, or private equity firm	**NO**
6. It has more than 50% owned by multiple business concerns that are VOCs, hedge funds, or private equity firms?	**NO**
7. The birth certificates, naturalization papers, or passports show that any individuals it relies upon to meet the eligibility requirements are U.S. citizens or permanent resident aliens in the United States.	**YES**
8. Is 50% or more of your firm owned or managed by a corporate entity?	**NO**
9. Is your firm affiliated as set forth in 13 CFR Section 121.103?	**NO**
10. It has met the performance benchmarks as listed by the SBA on their website as eligible to participate	**N/A**
11. Firms PI, CO, or owner, a faculty member or student of an institution of higher education	**NO**
12. The offeror qualifies as a: [ ] Socially and economically disadvantaged SBC [ ] Women-owned SBC [ ] HUBZone-owned SBC [ ] Veteran-owned SBC [ ] Service Disabled Veteran-owned SBC [**X**] None Listed	
13. Race of the offeror: [ ] American Indian or Alaska Native [ ] Native Hawaiian or Other Pacific Islander [ ] Asian [**X**] White [ ] Black or African American [ ] Do not wish to Provide	
14. Ethnicity of the offeror:	**NON-HISPANIC**
15. It is a corporation that has some unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have not been exhausted or have not lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability:	**FALSE**
16. Firm been convicted of a fraud-related crime involving SBIR and/or STTR funds or found civilly liable for a fraud-related violation involving federal funds:	**NO**
17. Firms Principal Investigator (PI) or Corporate Official (CO), or owner been convicted of a fraud-related crime involving SBIR and/or STTR funds or found civilly liable for a fraud-related violation involving federal funds:	**NO**

Signature:

Printed Name	Signature	Title	Business Name	Date
Ryan Wright	Ryan Wright	Chief Technology Officer	thatDot, Inc.	09/27/2023

Audit Information

Summary:

Has your Firm ever had a DCAA review?**NO**

VOL I - Proposal Summary

Summary:

Proposed Base Duration (in months):	**6**

Technical Abstract:

thatDot is a venture-backed American software company who created the revolutionary Quine streaming graph. The capabilities of this streaming graph, developed over multiple DARPA projects, provide a major head start toward solving the challenges of human interaction on dynamic knowledge graphs. With its “starnding query” capability, Quine helps humans interact with and understand dynamic knowledge graphs by monitoring the complete graph for user-defined patterns and triggering specified actions in real-time as the graph changes to fit the specified pattern. This ability produces results immediately and can be scaled to ingest millions of events per second.

With standing queries in a streaming graph as the revolutionary new state of the art, thatDot proposes a roadmap for 8 applications of this capability to address the Air Force’s need to equip users to interact with dynamically changing knowledge graphs. These proposed applications include: Machine-Assisted Graph Model Curation, Dynamic Schema Enforcement, Situational Awareness and Decision Making, Human-In-The-Loop Structure Learning, Pattern of Life Analysis, Threat Detection, and Graph Neural Networks for Similarity Measurement and Link Prediction.

Anticipated Benefits/Potential Commercial Applications of the Research or Development:

This proposal lays out a roadmap for 8 different applications of the new “Streaming Graph” capability to address the Air Force's need to allow a user to interact with a dynamic knowledge graph, make changes and additions to the knowledge graph, and suggest additional updates to surrounding nodes/edges in the graph. The applications proposed include: Machine-Assisted Graph Model Curation, Dynamic Schema Enforcement, Situational Awareness and Decision Making, Human-In-The-Loop Structure Learning, Pattern of Life Analysis, Threat Detection, and Graph Neural Networks for Similarity Measurement and Link Prediction.

Attention:

Disclaimer: For any purpose other than to evaluate the proposal, this data except proposal cover sheets shall not be disclosed outside the Government and shall not be duplicated, used or disclosed in whole or in part, provided that if a contract is awarded to this proposer as a result of or in connection with the submission of this data, the Government shall have the right to duplicate, use or disclose the data to the extent provided in the funding agreement. This restriction does not limit the Government's right to use information contained in the data if it is obtained from another source without restriction. This restriction does not apply to routine handling of proposals for administrative purposes by Government support contractors. The data subject to this restriction is contained on the pages of the proposal listed on the line below.

Addition:

Enter the page numbers separated by a space of the pages in the proposal that are considered proprietary:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

List a maximum of 8 Key Words or phrases, separated by commas, that describe the Project:

streaming, graph, data, artificial intelligence, real time, anomaly detection, novelty, knowledge

VOL I - Proposal Certification

Summary:

1. At a minimum, two thirds of the work in Phase I will be carried out by your small business as defined by 13 C.F.R Section 701-705. The numbers for this certification are derived from the budget template. To update these numbers, review and revise your budget data. If the minimum percentage of work numbers are not met, then a letter of explanation or written approval from the funding officer is required. **YES**

Please note that some components will not accept any deviation from the Percentage of Work (POW) minimum requirements. Please check your component instructions regarding the POW requirements.

Firm POW	**100%**
Subcontractor POW	**0%**

2. Is primary employment of the principal investigator with your firm as defined by 13 C.F.R Section 701-705? **YES**

3. During the performance of the contract, the research/research and development will be performed in the United States. **YES**

4. During the performance of the contract, the research/research and development will be performed at the offerors facilities by the offerors employees except as otherwise indicated in the technical proposal. **YES**

5. Do you plan to use Federal facilities, laboratories, or equipment? **NO**

6. The offeror understands and shall comply with export control regulations. **YES**

7. There will be ITAR/EAR data in this work and/or deliverables. **NO**

8. Has a proposal for essentially equivalent work been submitted to other US government agencies or DoD **NO**

components?	
9. Has a contract been awarded for any of the proposals listed above?	**NO**
10. Firm will notify the Federal agency immediately if all or a portion of the work authorized and funded under this proposal is subsequently funded by another Federal agency.	**YES**
11. Are you submitting assertions in accordance with DFARS 252.227-7017 Identification and assertions use, release, or disclosure restriction?	**NO**
12. Are you proposing research that utilizes human/animal subjects or a recombinant DNA as described in DoDI 3216.01, 32 C.F.R. Section 219, and National Institutes of Health Guidelines for Research Involving Recombinant DNA of the solicitation:	**NO**
13. In accordance with Federal Acquisition Regulation 4.2105, at the time of proposal submission, the required certification template, "Contractor Certification Regarding Provision of Prohibited Video Surveillance and Telecommunications Services and Equipment" will be completed, signed by an authorized company official, and included in Volume V: Supporting Documents of this proposal. NOTE: Failure to complete and submit the required certifications as a part of the proposal submission process may be cause for rejection of the proposal submission without evaluation.	**YES**
14. Are teaming partners or subcontractors proposed?	**NO**
15. Are you proposing to use foreign nationals as defined in 22 CFR 120.16 for work under the proposed effort?	**NO**
16. What percentage of the principal investigators total time will be on the project?	**20%**
17. Is the principal investigator socially/economically disadvantaged?	**NO**
18. Does your firm allow for the release of its contact information to Economic Development Organizations?	**YES**

VOL I - Contact Information

Principal Investigator

Name:	**Ryan Wright**
Phone:	**(810) 842-8368**
Email:	**ryan@thatdot.com**
Address:	**421 SW 6th Ave. Suite 300, Portland, OR 97204 - 1622**

Corporate Official

Name:	**Ryan Wright**
Phone:	**(810) 842-8368**
Email:	**ryan@thatdot.com**
Address:	**421 SW 6th Ave. Suite 300, Portland, OR 97204 - 1622**

Authorized Contract Negotiator

Name:	**Ryan Wright**

Phone:	**(810) 842-8368**
Email:	**ryan@thatdot.com**
Address:	**421 SW 6th Ave. Suite 300, Portland, OR 97204 - 1622**

Form Generated on 11/06/2024 02:38:11 AM

S.M.A.R.T. L.I.N.K.S.

"Streaming Modular Adaptation and Real-time Transformation for Learning and Inference in Knowledge Structures"

SBIR Topic: Interactive Knowledge Graphs for Situational Awareness

thatDot, Inc.

Table of Contents

Executive Summary

thatDot is a venture-backed American software company who created the revolutionary Quine streaming graph. The capabilities of this streaming graph, developed over multiple DARPA projects, provide a major head start toward solving the challenges of human interaction on dynamic knowledge graphs. With its “starnding query” capability, Quine helps humans interact with and understand dynamic knowledge graphs by monitoring the complete graph for user-defined patterns and triggering specified actions in real-time as the graph changes to fit the specified pattern. This ability produces results immediately and can be scaled to ingest millions of events per second.

With standing queries in a streaming graph as the revolutionary new state of the art, thatDot proposes a roadmap for 8 applications of this capability to address the Air Force’s need to equip users to interact with dynamically changing knowledge graphs. These proposed applications include: Machine-Assisted Graph Model Curation, Dynamic Schema Enforcement, Situational Awareness and Decision Making, Human-In-The-Loop Structure Learning, Pattern of Life Analysis, Threat Detection, and Graph Neural Networks for Similarity Measurement and Link Prediction.

Introduction

Knowledge graphs are a tremendously powerful tool to understand data, but technical limitations on speed and scalability have limited their application to small use cases that generally operate in "human-time." New technologies are opening the capabilities of these systems to unprecedented levels. Knowledge graphs are no longer static documents that rarely change but are becoming increasingly active and dynamic systems which show great promise for their role in modeling important systems to understand the world and live event taking place

However as these systems get faster, the ability for humans to understand has not kept pace. Tools were developed and staff was trained when graphs were slow and mostly static data repositories. For human understanding to keep up with the speed of technological development, new approaches are required.

thatDot is a venture-backed American software company known as the creators of the Quine streaming graph. A revolutionary technology developed over multiple DARPA programs[1] to solve previously impossible problems, Quine enables the understanding of high-volume streaming data at scale using event-driven graphs and the tools built on top of the graph. Initially designed to address the cybersecurity problem of detecting Advanced Persistent Threats (APTs), Quine has since seen applications in a wide collection of knowledge graph applications including: real-time attack path calculation for cyber threats at CrowdStrike, real-time risk scoring and asset allocation at JP Morgan Chase, following tainted funds through cryptocurrency blockchains, knowledge modeling and synthesis, data unification across enterprise data silos in data fabrics, and many more. In this proposal, thatDot lays out a

[1] Including "Transparent Computing": https://www.darpa.mil/program/transparent-computing and "ASKE": https://www.darpa.mil/program/automating-scientific-knowledge-extraction

research plan for building on top of the cutting edge streaming graph capabilities implemented in Quine and applying these capabilities to US Air Force applications.

Streaming Graphs

The New State of the Art

Quine

Over the last two decades, graphs have proven to be a highly effective way to organize data. Graph databases like Neo4j, AWS Neptune, and TigerGraph have grown to be important tools for representing data and transforming it into knowledge. Useful as these tools may be, their adoption has been severely limited by a fatal flaw: graph databases are slow and cannot scale to typical data volumes of the modern era. Even the fastest of these tools ends up being 3, 4, or 5 *orders of magnitude* too slow to be used in modern data pipelines.

thatDot created Quine[2] to solve this problem by building typical graph database functionality (e.g. storing and querying data with the SQL-like CYPHER graph query language) into an event-driven, actor based, event-sourced streaming graph. Quine creates a graph by representing each node with an Actor. The Actor Model was proposed by Carl Hewitt in the 1970s and it has seen a resurgence of use in the last decade as the ideal model for high-volume event-driven data processing pipelines. Quine integrates Actors deep into the graph model by using one or more actors to represent every node in the graph. Actors are lightweight processes that contain internal state and communicate exclusively through asynchronous message passing. An Actor has a mailbox where it receives messages, and

[2] See: https://quine.io

when scheduled on a CPU, it processes those messages from its mailbox in a single-threaded fashion. An actor can perform arbitrary computation upon receiving a message, which allows it to perform all the duties required of a node in a graph. The Actor foundation allows nodes in a Quine graph to efficiently compute query results, monitor themselves and their neighbors for matching user-defined patterns, and trigger other actions when they witness certain kinds of events. Actors can also be distributed over a very large cluster of computers to scale horizontally to support any size of workload.

With Actors at the heart of the graph in Quine, we employ other techniques to compliment the new functionality they afford. Data is saved to disk using a technique called "event-sourcing," where only the *changes* to each node are actually saved to disk, and then added very efficiently to an append-only log. This makes write operations very fast and the ideal way to support stream processing while ensuring updates are durably stored on disk. With a history of changes for each node, the graph is also fully versioned, and the history of the full graph can be queried by simply including a timestamp with the user's Cypher query. With this capability, Quine enables querying all historical moments in the past to see what the data used to be.

Standing Queries are a powerful feature enabled by Quine's unique design. These are the heart of how the SMART LINKS system can synthesize and understand new data ingested, and make timely interactive recommendations to human users. A Standing Query monitors the graph while it is changing for user-specified patterns of any complexity. When the data changes in a way so as to match the desired pattern, it triggers a Standing Query action. These actions occur in real-time as the data changes. They can alert analysts, publish data to downstream systems (like thatDot Novelty), or call back into the graph to create or update elements in the next tier of understanding in a new layer of the graph.

thatDot has been told that the Air Force is already using Quine as a part of the “BDP” or “Big Data Platform” in the ELICSAR program.[3] As an optional component in that platform, the ability to bring streaming knowledge synthesis and real-time pattern matching can already deliver tremendous benefits. Extending that capability through the work proposed on the SMART LINKS project herein then has the potential to multiply the impact to other users inside the Air Force and across the DoD.

Novelty Detection

Novelty Detection is a new class of anomaly detection tool built on top of the Quine streaming graph. Virtually all other anomaly detection tools require only numbers as their input data. But most behavioral data (of humans and computer systems) isn’t numeric but rather is categorical. Categorical data is just all the data which is not a number. Values like email addresses, username, file paths, IP addresses, and even ciphertexts are all categorical values. Using categorical values in traditional anomaly detection is virtually impossible. While data scientists do have tools to represent categorical values as numbers (e.g. one-hot encoding, binary encoding, etc.), doing so increases the vector dimensions enormously! One-hot encoding (the most common approach) creates a new dimension for every single new *value* represented in the dataset; so every time a field has a new value, a new dimension must be added to the vector. When performing anomaly detection, the practitioner faces the “curse of dimensionality”: as the number of dimensions increases, soon everything becomes an

[3] https://elicsar.cce.af.mil/

anomaly. This leads to a huge number of false positive results. To make matters worse, these approaches cannot handle values they haven't seen before, making them utterly unusable for streaming data applications.

thatDot's Novelty Detector takes a new approach. It uses the Quine streaming graph under the hood to build a graph representation of the categorical data that streams through. With a graph representation, Novelty can analyze each path through that graph to understand how common that path is, what alternatives there have been and how common they are, and ultimately use conditional probability in an information theoretic analysis of those graph paths to measure how novel each observation is compared to all observations which have streamed in previously. This allows Novelty Detector to perform a self-training unsupervised and streaming/online learning model to measure how unusual or "truly novel" each event is, rather than relying on heuristics or approximations. Novelty Detector processes streaming data in real time to rank each event by how unusual it is—even if it has never been seen before.

Unsupervised learning allows for better detection of the unknown unknowns to stop cyber threats. Since supervised learning relies on labeled datasets with known patterns, it struggles to identify new or previously unseen threats. Unsupervised learning, on the other hand, can detect anomalies and novel patterns in data without prior knowledge of what constitutes a threat. Malicious actors are dynamic and continuously evolving to evade detection mechanisms. Novelty Detection automatically learns the fingerprint of normal behavior, automatically adapting to continuously learn from new data and identify and flag deviations from normal behavior. This behavioral monitoring approach fundamentally changes the cat-and-mouse game that cyber defenders typically face, which is crucial for detecting evolving cyber threat tactics.

Applications

At the heart of thatDot's proposal to the Air Force for the "Interactive Knowledge Graphs for Situational Awareness" program is an important new capability: to know automatically ***WHEN*** important relationships and data structures have appeared (or disappeared) in the ever-growing graph model, and then to trigger action with that data automatically and in real-time. That action can: a.) prompt a human to make a decision, b.) feed an external automated system in real-time, or even c.) feed back into the graph to update data or compute the next step in an algorithm to provide richer answers to the people who depend on *understanding* that data.

Herein we lay out a roadmap of capabilities to be explored throughout the phases of this program—each of them enabled by the unique capabilities of the Quine streaming graph where the fundamental technologies have already been proven, and with the team of experts who work to deliver these capabilities daily.

1: Machine-Assisted Graph Model Curation

Standing queries are a revolutionary new capability to come to knowledge graph systems. With a standing query, a user can express a pattern of interest and rely on the system to inform them every time a new result matches the pattern they are interested in. Those patterns can be positive patterns (matching a specific structure) or negative patterns (matching a base pattern, but failing to match an extension of that base pattern). This makes standing queries an ideal tool for a user to monitor a graph while it continues to change and ensure that various characteristics the user expects of the data continue to hold. Any variation of the data from what the user has expressed is delivered as results that can link to the specific data.

We propose creating an interface to enable a user to define a new standing query from their visual interaction with the underlying data in the knowledge graph. The current capability requires a user to leave their graphical environment and make an API call to express a new Cypher query and output handler as their new Cypher query. This process is cumbersome and requires skills beyond those needed for visual analysis and interaction.

With this proposed interface, a user can explore the graph using Quine's built-in Exploration UI[4], and upon finding a surprising pattern (because of unexpected extant graph structure or missing graph structure) the user could easily define a new standing query from the visualization they are already looking at. The new standing query would allow the user to: a.) be notified if the pattern occurs again, b.) define a mitigation to adjust the data to its desired form, c.) publish the unexpected data to downstream or external systems, or d.) update an algorithm running on the graph, or e.) trigger any other action programmatically.

2: Dynamic Schema Enforcement

One of the great advantages of a graph data model is that it is very flexible. However that advantage also becomes a disadvantage in many applications. A graph model is considered a "NoSQL" data model because it does not require or enforce a schema definition for data to be written in. The downside of this model is that: a.) it is hard for a new query writer

[4] https://quine.io/getting-started/exploration-ui/

to know what is worth querying without a structure of the data to start from; and b.) the data might change in unexpected ways, altering the results they expect to see.

We propose a new capability for an end user to create a “dynamic schema” on their knowledge graph. A dynamic schema is the ability—at any point during the operations on a graph—to specify constraints that must hold on the graph structure, or a query mutation will fail and return an error. A dynamic schema is not possible or practical on other graph systems, but because nodes in a Quine graph are backed by actors, and those actor can perform arbitrary computation during other graph operations, enforcing a dynamic schema is within reach.

A dynamic schema can be thought of as a standing query which, upon matching a relevant node, will actively check subsequent operations on that node and cause incompatible types of operations to fail with a “Schema Violation” error. In order to be applicable, a node must first meet the criteria for the schema constraint—which can be as simple or complex as desired. Once it meets that criteria, the specified types of operations will be disallowed on that node.

As a concrete example: A node in the graph may represent a user account, as signified by the presence of a `userId` property on that node. User accounts may be granted access to other systems (signified by a `systemId` property) by creating a `can_access` edge from the user account node to the system node. A dynamic schema constraint could be added to the system node to ensure at least one user will always have access. To enforce this, a dynamic schema constraint can be set on the system node which applies whenever the system has one or more incoming `can_acccess` edges (so the schema is only enforced after the first user has been granted access). If a user account is about to be deleted or have its `can_access` edge removed, and that user is the last user connected with a `can_access` edge to that system node, then the operation attempting to remove the user returns an error with an explanation that a dynamic schema violation has occurred and explaining the requirement to ensure at least one user will have access to the system in question.

The formal theory behind dynamic schemas is based on “structural types”[5] from type theory, which is also the theoretical model underlying standing queries.

3: Situational Awareness and Decision Making

Quine is an event-driven graph, which means it reacts to changes. Any event that causes data to change anywhere in the graph has the potential to perform arbitrary computation or to trigger other events (inside or outside of the system). Having an event-driven capability on a graph makes it the perfect infrastructure tool to synthesize data from many remote sensors, troops, vehicles, or other assets. Modeling those assets in a graph becomes the foundation for a highly adaptive and useful visualization to provide situational awareness for highly complex scenarios unfolding in real-time.

We propose building a graph model for situational awareness and a UI integration which uses Quine’s event-driven capability to update a visualization (e.g. a battlefield map) in real-time with the results of data reported from remote sensors. More than simply moving the positions of icons over a map, to effectively visualize a large number of data points requires aggregating and interpreting groups of data points depending on the circumstances of how the data changes over time and over a complex set of other characteristics (e.g. in relation to geographical features, progress toward mission goals, etc.).

[5] https://en.wikipedia.org/wiki/Structural_type_system

Our proposal would build multiple graph layers, where each successive layer is an interpretation of the layer below it. As changes ripple through each layer of the graph, their updates cause the layer above to update in real-time as well. Each graph layer shows a smaller amount of total data than the layer below it that it interprets, but information in the higher layer is more abstract than the layer below it. While the lowest layer would include every data point for every sensor reporting, the top layer would simply sum up the total progress as being “on target” or not. As you move from the bottom to the top layer, each successive layer contains less total data (number of nodes), but each node represents an increasingly rich concept.

4: HITL Structure Learning

The richness of graph structure is on display when ingesting simple events which, once connected, can yield important and complex relationships among the paths between them. The structure of the data written in is often not the determinant of what patterns are valuable to query out. Sometimes a creative analyst can imagine which patterns are worth querying out of the emergent graph structure, but there is almost always a collection of interesting patterns that would be useful to know which the analyst cannot conceive of querying.

We propose a tool for learning those important patterns from the graph through an automated human-in-the-loop (HITL) process of structure learning. Sometimes called “pattern learning” or “query learning,” the process of structure learning on a graph begins with the data in the graph, and ends with a pattern or a query that describes the subsections (or subgraphs) in the data which are useful to query for.

This process would begin by using machine learning tools to measure the similarity of nodes in the graph; see the section of Graph Neural Networks (GNNs) below. Nodes that are measured to have a high similarity will have their neighboring structure parsed to determine small scale structure found in common among similar nodes. The small-scale parsed structures will be presented to a human for evaluation and—crucially—for naming. Any pattern approved by a human evaluator and able to be named will be fed back in as input to repeat the process. With the named structure as input, the process will repeat, but this time using the named structure to rewrite the graph wherever that structure occurs into a single node. Simplifying the graph structure based on the named pattern gives us a new iteration of the graph structure to measure for similarity, parse out common structure, and again present the newly found common structures to a human. Upon this second interaction, new patterns presented to the human analyst to approve and name will be composed of instances of their prior simpler named structure. This process continues for as long as the human is able to identify and name useful structures.[6]

5: Pattern of Life Analysis

Graphs are a useful tool for understanding behavioral data of subjects. These subjects can be humans, computer systems, artificially intelligent agents, or more. Recording a series of behavioral observations naturally draws a graph as subjects interact with each other and objects in the world. Understanding commonalities in this space can follow the same general pattern as described above in the “Pattern Learning” section to find *types of behavior* that are

[6] This proposal is based on prior led by the same principal investigator, Ryan Wright, which showed this approach to be effective: https://arxiv.org/abs/1908.02947

common but useful to understand. However, much of the value in pattern of life analysis is in recognizing *unusual* patterns in the behavior of subjects.

We propose a tool for real-time ranking of how unusual behavioral patterns are that occur in a graph. This tool would be able to a.) assemble individual behavioral events into a graph in order to b.) connect them to relevant context and c.) make simple measurements, before d.) sending complete observations to the Novelty Detection component to learn a thoroughgoing model of contextually aware behavioral patterns, and e.) scoring them in real-time to understand how novel that behavior is compared to the normal behavior pattern for each individual. To develop this tool would be a matter of using Quine and Novelty Detection together and building an integration fit to data that corresponds to a real-world Air Force use case.

Two key factors are critical for this approach: 1.) The learning model is entirely unsupervised. No training data is required. 2.) The model that is learned is contextually aware and able to distinguish between events that are simply new and previously unseen vs. truely novel. This approach has proven very effective in pattern of life analysis for cybersecurity to measure in real-time the suspicious behaviors that call for swift action to stop a cyber threat.

6: Threat Detection

Interactive real-time graphs which can connect individual log events and measure the novelty of the events they observe are emerging as a critical capability for threat detection and cybersecurity. Connecting events gives a human analyst the context they need to understand the events. But connecting data fast enough to keep up with non-trivial data volumes has historically been a challenge for graph systems. Quine has the ability to scale horizontally and has been demonstrated ingesting data over 1,000,000 events per second, while simultaneously finding and publishing 21,000 standing query results every second. This scale represents an improvement of 1,000-100,000 times beyond any other graph system. It opens up deep graph analysis to the data volumes needed to analyze and understand cybersecurity logs at the scale of large organizations. When combined with Novelty Detection, we have the capability to aid a human in prioritizing their analysis to ensure that emerging threats are properly identified in time to stop the threat.

We propose an application which combines the Quine streaming graph with Novelty Detection to find and prioritize emerging threats based on the novelty of their behavioral patterns. This capability is resilient to new and emerging threats threats because it evaluates the *behavior* of system *in context* and delivers a.) timely alerts to human analysts with b.) a link to interactively explore the relevant graph structure and c.) any surrounding context the analyst needs to understand the threat, and d.) a ranking of how novel this threat is compared to all other activity in the system. Demonstrating this capability will require either accessing or generating a proper dataset to simulate real-world background traffic, realistic malicious activity, and the integration of two systems to consume and evaluate the data.

Our team has deep experience performing this kind of demonstration and deployments in real-world environments. The US Navy is currently exploring such a deployment for detecting malicious adversaries onboard Navy ships in a DDIL environment. In the commercial sector, Crowdstrike has deployed our tools to perform real-time analysis of potential attack paths from compromised machines to critical assets. Faster threat detection enables a critical difference in time to respond and success at stopping the threat.

7: GNNs for Similarity Measurement and Link Prediction

Graph Neural Networks (GNNs) are proving to be the powerful and necessary complement to LLM applications. While LLMs have amazed with the quality of their English responses, research has shown they are especially bad at logical reasoning[7] and graph-oriented tasks in particular.[8] However, graph neural networks are proving to provide the kind of graph-based reasoning for which LLMs are ill-suited.[9]

We propose a toolset for iteratively exploring updates to graph structure derived from GNNs reasoning over graphs. In particular, the tool-enabled workflow would allow a user to automatically train a neural network over graph data which can then propose back to the user two kinds of suggestions: 1.) nodes or graph structures that are similar to the nodes they are investigating, and 2.) predicted edges missing from the dataset. Multiple GNNs techniques will be explored and their tradeoffs evaluated, including: Node2Vec, GraphSAGE, Graph Convolutional Networks, and Graph Attention Networks. Each of these techniques serves a different purpose, and includes varying tradeoffs.

A key challenge to operationalizing GNNs in practice is that the underlying graph data they need for training isn't naturally static, it's continually updating. Most datasets continue to change as time marches on. Training neural networks can be time consuming when the dataset is large, and this exacerbates the problems in any approach that needs to hold the data

[7] See recent research from Apple: https://arxiv.org/abs/2410.05229
[8] See recent research from Google researcher Bryan Perozzi and team: https://arxiv.org/abs/2310.04560
[9] See our related work here: https://www.thatdot.com/blog/graph-neural-networks-for-quine/

constant—often for days at a time—while the neural network data is prepared and trained. One of the key capabilities of Quine that we'll explore in this area is the ability to query the historical state of the graph even while it continues to change. Quine's internal data model produces a fully-versioned graph where every single change is saved individually and conveniently accessible. That means that queries can be run on historical data in exactly the same way that they are run on present data, just by including a timestamp with the query. That gives a dynamically changing graph the ability to be queried at a previous fixed state (including queries that dynamically generate random walks).

Phase 1 Statement of Work

Scope

This proposal will research techniques that enhance a user's ability to interact with dynamic knowledge graphs. In earlier sections, we laid out a roadmap of research topics which we expect will all provide value to the Air Force mission behind this program. For execution during Phase 1, we propose tackling the first option listed under "Applications" above: Machine-Assisted Graph Model Curation. Remaining topics can be explored in later phases of this work.

However, our team is interested in maximizing the value to the Air Force in this research. Therefore, if a different order of execution would be most appropriate, our team is willing to adapt the plan as laid out above to fit. To that end, our tasks, milestones, and deliverables incorporate discussion and feedback to adapt the work to the topics most relevant to the Air Force.

Task Outline

1. Coordinate technical use case with customer
2. Identify use case dataset
3. Build streaming graph implementation
4. Report on results.

Milestone Schedule

ID	Title	Description	Delivery
01	Kickoff Meeting	Kickoff meeting and associated materials	30 days from award
02	Research Plan & Progress Report	Research plan for using Quine and Novelty Detector to address the goals for interactive knowledge graphs.	60 days from award
03	Identified Dataset & Progress Report	Contractor will deliver confirmation of identified datasets suitable for the research plan, and a status report.	90 days from award
04	Interim Demo	Contractor will demonstrate preliminary capabilities for agreed upon experiments and datasets via teleconference	120 days from award
05	Final Demo	Final demonstration of capability via teleconference	180 days from award
06	Final Report with SF 298	Report on the final work performed during Phase 1 and results achieved	180 days from award

Deliverables

- Kickoff Meeting materials: Meeting minutes and
- Research Plan: A report of the updated research plan incorporating customer feedback

- Dataset Confirmation: An explanation of the identified dataset which matches the customer's desired use case application of the research
- Interim Demo: A demonstration of interim progress made toward the overall use case implementation.
- Final Demo: A demonstration of the final progress made during Phase 1 toward the overall use case implementation.
- Final Report with SF 298: A written summary of the results accomplished during phase 1 and a standard form 298.

For the code developed on this program, we plan to deliver that code into the open source project that is published at https://quine.io with code freely available from https://github.com/thatdot/quine The Air Force and the DoD at large will be able to freely use all the code artifacts produced as a result of this proposal.

Progress Reports

Our desire is to make this research maximally effective for the Air Force. Therefore, we incorporate into our proposal the ability to adapt to specific customer needs and integrate customer direction into our work. We will report on the state of research as it adapts to customer needs through periodic status reporting. The delivery of status reports will align with the timeframes described in the Milestone Schedule above.

Final Report with SF 298

At the completion of Phase 1, the contractor will submit a final report to summarize the work done to date, including key performance metrics (including accuracy and completeness) and lay out directions for future continuation of the work performed. This report will include a Standard Form 298.

Related Work

From 2015 to 2019, Ryan Wright led a team of researchers from Galois, PARC, the University of Oregon, and the University of Edinburgh on the DARPA Transparent Computing program. This program, organized by DARPA program manager Dr. Angelos D. Keromytis, was funded specifically to accomplish the 6-1 and 6-2 research required to detect highly sophisticated, nation-state-level attackers who are already inside of computer systems. Research on this program led to the two key developments described in this proposal: the Quine streaming graph and Novelty Detection. With these two revolutionary technologies, the team led by Wright demonstrated automated detection of Advanced Persistent Threats (APTs) by the end of the program. Wright founded thatDot to commercialize these technologies and bring them to the field as deployable commercial applications. In recent months, thatDot has worked with the US Navy Undersea Naval Warfare Office (with Ian T. Russell and Kathleen Hourihan), the Office of Naval Research (with Dr. Ryan Craven), AFWERX, and Lockheed Martin (with Jake Wertz, Scott Timme, William Moldenhauer, Char Carrie, and Tyler William) to successfully test applications of these technologies to protect systems onboard aircraft, ships, missile systems, and submarines.

Future Research and Development

Future research and development objectives are focused on the combination of Quine Enterprise and Novelty Detector into a cohesive package that is forward deployable in a containerized package. These packages will enable the organization to quickly deploy and maintain the capability, ingest large volumes of data and impact decision making without delays. Future development objectives will also focus on scalability to ensure that any platform across the Navy can interoperate with other instances of this capability or with legacy tools.

Key Personnel

thatDot is proud to introduce its key personnel, a team of highly qualified and accomplished individuals who bring a wealth of expertise and experience to our organization. Committed to excellence and innovation, our key personnel play a pivotal role in driving the success of thatDot and ensuring our continued growth in the ever-evolving landscape of technology and data analytics. With their diverse skill sets and unwavering dedication, they embody our company's values and are instrumental in shaping our vision for the future.

KEY PERSONNEL SUMMARY

Name	**Title**	**Employer**	**Description**	**US Citizen**
Ryan Wright	CTO	thatDot	Ryan Wright is the creator of Quine (https://quine.io) and has been leading software teams focused on data infrastructure and data science for two decades. He was principal engineer, architect, and director of engineering at several software startups. While at	Yes

			Galois for 6 years, Ryan served as principal investigator and more on DARPA-funded research programs such as Transparent Computing, CFAR, ASKE, Safeware, Brandeis, and others. He has founded four different companies, and is currently the founder and CTO of thatDot—the company supporting the new open source streaming graph, Quine.	

Foreign Citizens

thatDot does not have any foreign citizens, U.S. permanent residents, or individuals with dual citizenship employed in any capacity within our team for this project. Therefore, there are no individuals to report in terms of their country of origin, visa or work permit status, or anticipated level of involvement on this project. Our team is composed entirely of individuals who do not fall into this category, ensuring full compliance with all project requirements and regulations.

DSIP Defense SBIR/STTR Innovation Portal

Cost Volume

SBIR Phase I Proposal

Proposal Number	F244-0001-0097
Topic Number	AF244-0001
Proposal Title	S.M.A.R.T. L.I.N.K.S. - "Streaming Modular Adaptation and Real-time Transformation for Learning and Inference in Knowledge Structures"
Date Submitted	11/06/2024 02:38:08 AM

Firm Information

Firm Name	thatDot, Inc.
Mail Address	421 SW 6th Ave. Suite 300, Portland, Oregon, 97204
Website Address	https://www.thatdot.com
UEI	YDL7H2F7BH93
Cage	8EQU1

Total Dollar Amount for this Proposal	$140,000.00
Base Year	$140,000.00
Year 2	$0.00
Technical and Business Assistance(TABA)- Base	$0.00
TABA- Year 2	$0.00

Base Year Summary

Total Direct Labor (TDL)	$140,000.00
Total Direct Material Costs (TDM)	$0.00
Total Direct Supplies Costs (TDS)	$0.00
Total Direct Equipment Costs (TDE)	$0.00
Total Direct Travel Costs (TDT)	$0.00
Total Other Direct Costs (TODC)	$0.00
G&A (rate 0%) x Base ()	$0.00
Total Firm Costs	$140,000.00
Subcontractor Costs	
Total Subcontractor Costs (TSC)	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$140,000.00
TABA	$0.00

Year 2 Summary

Total Direct Labor (TDL)	$0.00
Total Direct Material Costs (TDM)	$0.00
Total Direct Supplies Costs (TDS)	$0.00

Total Direct Equipment Costs (TDE)	$0.00
Total Direct Travel Costs (TDT)	$0.00
Total Other Direct Costs (TODC)	$0.00
G&A (rate 0%) x Base ()	$0.00
Total Firm Costs	$0.00
Subcontractor Costs	
Total Subcontractor Costs (TSC)	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$0.00
TABA	$0.00

Base Year

Direct Labor Costs

Category / Individual-TR	**Rate/Hour**	**Estimated Hours**	**Fringe Rate (%)**	**Fringe Cost**	**Cost**
Computer and Information Research Scientist/ Principal Investigator	$250.00	208			$52,000.00
Software Developer/ Engineer	$160.00	550			$88,000.00

Subtotal Direct Labor (DL)	$140,000.00
Labor Overhead Cost	$0.00
Total Direct Labor (TDL)	**$140,000.00**

G&A (rate 0%) x Base ()	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$140,000.00
TABA	$0.00

Year 2

Direct Labor Costs

Category / Individual-TR	**Rate/Hour**	**Estimated Hours**	**Fringe Rate (%)**	**Fringe Cost**	**Cost**
Computer and Information Research Scientist/ Principal Investigator	$250.00	0			$0.00

Subtotal Direct Labor (DL)	$0.00
Labor Overhead Cost	$0.00
Total Direct Labor (TDL)	**$0.00**

G&A (rate 0%) x Base ()	$0.00
Cost Sharing	-$0.00
Profit Rate (0%)	$0.00
Total Estimated Cost	$0.00
TABA	$0.00

Explanatory Material Relating to the Cost Volume

The Official From the Firm that is responsible for the cost breakdown

Name: Ryan Wright
Phone: (810) 842-8368
Phone: ryan@thatdot.com
Title: Proposal Owner

If the Defence Contracting Audit Agency has performed a review of your projects within the past 12 months, please provide: No

Select the Type of Payment Desired: Partial payments

Cost Volume Details

Direct Labor

Base

Category	Description	Education	Yrs Experience	Hours	Rate	Fringe Rate	Total
Computer and Information Research Scientist	Principal Investigator	Master's Degree	22	208	$250.00		$52,000.00
Software Developer	Engineer	Bachelor's Degree	10	550	$160.00		$88,000.00

Are the labor rates detailed below fully loaded? **YES**

Please explain any costs that apply.
Standard burdened costs including benefits, health insurance, supplies, etc.

Provide any additional information and cost support data related to the nature of the direct labor detailed above.
Rates compare favorably.

Direct Labor Cost ($): $140,000.00

Year2

Category	Description	Education	Yrs Experience	Hours	Rate	Fringe Rate	Total
Computer and Information Research Scientist	Principal Investigator	Master's Degree	22	0	$250.00		$0.00

Are the labor rates detailed below fully loaded? **YES**

Please explain any costs that apply.
Same as previous year.

Provide any additional information and cost support data related to the nature of the direct labor detailed above.
Rates compare favorably.

Direct Labor Cost ($): $0.00

Sum of all Direct Labor Costs is($):	$140,000.00

Overhead

Base

Labor Cost Overhead Cost	**$0.00**
Overhead Comments:	
Overhead Cost ($):	**$0.00**

Year2

Labor Cost Overhead Cost	**$0.00**
Overhead Comments:	
Overhead Cost ($):	**$0.00**
Sum of all Overhead Costs is ($):	**$0.00**

General and Administration Cost

Base

G&A Rate (%):	**0**
Apply G&A Rate to Overhead Costs?	**NO**
Apply G&A Rate to Direct Labor Costs?	**NO**
Please specify the different cost sources below from which your company's General and Administrative costs are calculated.	
G&A Cost ($):	**$0.00**

Year2

G&A Rate (%):	**0**
Apply G&A Rate to Overhead Costs?	**NO**
Apply G&A Rate to Direct Labor Costs?	**NO**

Please specify the different cost sources below from which your company's General and Administrative costs are calculated.

G&A Cost ($):	**$0.00**
Sum of all G&A Costs is ($):	**$0.00**

Profit Rate/Cost Sharing

Base

Cost Sharing ($):	**-$0.00**
Cost Sharing Explanation:	
Profit Rate (%):	**0**
Profit Explanation:	
Total Profit Cost ($):	**$0.00**

Year2

Cost Sharing ($):	**-$0.00**
Cost Sharing Explanation:	
Profit Rate (%):	**0**
Profit Explanation:	
Total Profit Cost ($):	**$0.00**
Total Proposed Amount ($):	**$140,000.00**

CERTIFICATE OF COMPLETION

THIS CERTIFICATE IS PRESENTED TO

Ryan Wright, thatDot, Inc.

FOR SUCCESSFULLY COMPLETING FRAUD, WASTE AND ABUSE TRAINING AND MEETING ALL REQUIREMENTS SET FORTH BY THE OFFICE OF SMALL BUSINESS PROGRAMS

DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA

Nov 04, 2024	Nov 04, 2025
COMPLETION DATE	EXPIRATION DATE