

UDC 004.056

A Cognitive Cyber Range Based on Two-Sided Semantic Network Construction

Maksym Osinnii¹, Dmytro Lande¹, Denys Honcharov¹, Mykyta Lytvyn¹¹ National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine

Abstract

Scenario analysis in a cyber range requires a structured representation of infrastructure and threat knowledge. This paper proposes a cognitive cyber range model based on two-sided semantic network construction and examines its graph component using a software prototype. The theoretical model combines forward and backward expansion, scenario synthesis, and multicriteria assessment. The prototype represents document-level concept co-occurrence as an undirected contextual semantic network, weights its links using normalized pointwise mutual information, and generates ranked paths for subsequent interpretation by a large language model. Three search configurations are evaluated on 20 test scenarios of cyber intrusions into infrastructure. Mean reference-step recall across up to five paths is 0.6083 for bidirectional search performed separately for each supplied initial concept and 0.2042 for one-sided search without supplied initial concepts. These values compare configurations that differ in both input information and traversal procedure; they do not isolate the effect of the search algorithm. Recall of the first-ranked path, preservation of step order, and node precision are assessed separately. Literal identifier grounding before registry validation supports the preparation of source documents. The prototype supports traceable exploration of scenario hypotheses. Causal consistency, action feasibility, numerical multicriteria assessment, and synchronization with live infrastructure remain outside the validated implementation and require further research.

Keywords: cognitive cyber range, semantic digital twin, contextual semantic network, bidirectional search, scenario analysis

Introduction

Cyber ranges provide controlled environments for reproducing cyberattacks, investigating vulnerabilities, testing security controls, and training personnel. Experiments can be conducted on modeled infrastructure without exposing operational information systems to risk [1]. We examine the analytical task that precedes scenario selection: constructing a space of alternatives from a contextual semantic network and explaining the relationships it contains.

A semantic digital twin represents information infrastructure together with knowledge of threats and the current information context. Here, we investigate the document-derived graph component of this model. Statistical associations between concepts are used to generate hypotheses for scenario analysis. Connecting the model to live infrastructure and representing the preconditions

for defensive actions remain tasks for further development.

Advances in artificial intelligence, semantic networks, digital twins, large language models (LLMs), and retrieval-augmented generation (RAG) create new opportunities for cyber threat analysis. A language model generates a textual response and does not, by itself, guarantee that the alternatives considered are comprehensive, that the resulting scenario is causally consistent, or that it reflects the actual state of an information system.

1. Related Work

Zhang et al. [2] review the current state and future directions of cyber ranges. VSDL [3] generates virtual scenarios from formal specifications with constraint checking. These studies establish requirements for scenario automation and consistency with a specified infrastructure. The present study represents

possible scenarios through relationships between concepts.

An ontology-based cyber range model [4] defines domain concepts and relationships and supports logical validation. Knowledge graphs combined with language models [5] are used to extract cybersecurity concepts and identify relationships between them. The contextual semantic network examined here is based on concept co-occurrence in documents. These statistical associations provide a basis for searching for hypotheses whose meaning is subsequently assessed by an analyst.

The combination of semantic networks and LLMs is discussed in [6], and a digital twin model for a cyber range is presented in [7]. The methodological foundations include systems analysis [8] and semantic networks for open-source intelligence (OSINT) [9]. LLM-Twin [10] investigates network digital twins, while Nguyen [11] examines an ontological representation of humans for cybersecurity simulations. These approaches motivate the use of structured knowledge to interpret system states and analyze possible actions.

The capabilities and vulnerabilities of large language models are reviewed in [12]. Their application to threat analysis and decision support is discussed in [13], the evaluation of OSINT chatbots in [14], and the use of RAG for cyberattack investigation in [15]. In the proposed approach, the language model explains alternatives previously generated by a graph algorithm. The retrieved paths and source documents form its context.

Multicriteria scenario selection requires the reconciliation of assessments of feasibility, risk, and resource availability. The agent-based approach [16] involves distributed assessment and decision coordination. Study [17] applies iterative MRRW-PageRank link weighting to penetration scenarios. The foundations of multi-agent reinforcement learning are presented in [18], and its application challenges are discussed in [19]. The theoretical formulation in this study provides for assessment by specialized agents; the software prototype implements role-based document analysis without agent reinforcement learning.

The reviewed studies address scenario formalization, semantic knowledge representation, and language-based interpretation. For the proposed model, an open question is how well search in a concept co-occurrence network reproduces reference attack-

step sequences and how the configuration of initial information and traversal affects the result. A further distinction is required between the correspondence of a single path to the reference and the combined matches found across several alternatives.

2. Problem Statement and Research Aim

The aim of this study is to develop a cognitive cyber range model based on two-sided semantic network construction and experimentally evaluate its graph component. This requires defining the knowledge representation and link characteristics, formalizing scenario synthesis and assessment, implementing the search for alternatives, and comparing the results with reference attack-step sequences.

The information basis comprises documents describing the state of infrastructure components, cyber threats, and vulnerabilities. A task specifies a goal concept and, where needed, initial concepts selected by an analyst. The objective is to generate a set of paths associated with these concepts, rank the alternatives, and present them together with link characteristics, sources, and explanations.

In the graph formulation, the goal corresponds to a specified concept, and a retrieved path is a sequence of related concepts. As the underlying network is undirected, the direction in which a path is presented determines its reading order. Interpreting the path as an action scenario additionally requires establishing the meaning of its transitions and the preconditions for performing them.

The vertices of the contextual semantic network represent domain terms, vulnerability identifiers from the Common Vulnerabilities and Exposures (CVE) system, and technique identifiers from the MITRE ATT&CK knowledge base. Assets, resources, and security controls are represented by document terms. Further formalization requires defining entity types and the relationships between them.

A cyber defense scenario should specify the preconditions and effects of actions, their execution order, and their admissibility under resource, time, and regulatory constraints. This study computationally investigates the generation of graph-based hypotheses that an analyst can use to refine a scenario and formulate subsequent tasks.

Multicriteria assessment should account for technical feasibility, risk, implementation time, required resources, regulatory compliance, and the competencies of personnel. Its formal formulation is presented in Section 3.3. The experimental part addresses the structural ranking of paths and their correspondence to reference attack-step sequences.

The contribution of this study is to specify a semantic digital twin model for algorithmic generation of alternatives and experimentally compare three search configurations on a common set of test scenarios. The evaluation separately measures correspondence between the first path and the reference, preservation of step order, and recall across the set of retrieved paths. It identifies cases in which reference concepts occur in different alternatives, while an individual path reproduces only part of the sequence.

3. Proposed Model and Methods

3.1. Contextual Semantic Network

A semantic network is used as the principal knowledge representation. In the implemented prototype, a contextual semantic network represents concepts and their co-occurrence in documents as an undirected graph:

$$G = (V, E), \quad (1)$$

Network vertices correspond to concepts extracted from documents, and edges represent their co-occurrence. Link attributes include co-occurrence frequency and references to the corresponding sources. Standardized CVE and MITRE ATT&CK identifiers are represented by anchor nodes, which are taken into account when assigning link weights.

Let N denote the number of documents included, $f(u)$ the number of documents containing concept u , and $c(u, v)$ the number containing both concepts. Each document contributes at most one to each corresponding count. For positive frequencies and a nonempty corpus, the probabilities are defined as follows:

$$\begin{aligned} p(u) &= \frac{f(u)}{N}; \\ p(u, v) &= \frac{c(u, v)}{N} \end{aligned} \quad (2)$$

For a pair whose joint probability is strictly between zero and one, normalized pointwise mutual information (NPMI) is defined by (3),

where $\ln$ denotes the natural logarithm. This measure quantifies the statistical association between concepts in the corpus [20].

$$I(u, v) = \frac{\ln\left(\frac{p(u, v)}{p(u)p(v)}\right)}{-\ln p(u, v)} \quad (3)$$

The measure I characterizes the strength of the statistical association between concepts. In boundary cases, I is computed using the rules described in Section 3.5.1. The symbol ε denotes a small positive constant that ensures numerical stability.

Link characteristics account for statistical similarity between concepts and whether they are standardized identifiers. The coefficient b specifies the boost assigned to a node. The effective NPMI-based weight w is defined by (4), and its frequency-based alternative by (5). These weights are used for path search in the network.

$$w(u, v) = \max(1 + I(u, v), \varepsilon) \times \sqrt{b(u)b(v)} \quad (4)$$

$$w(u, v) = c(u, v)\sqrt{b(u)b(v)} \quad (5)$$

The positive edge cost d is defined by (6). Let P be a path with $h(P)$ edges and let $v(i)$ denote the node at position i . In (7), $\sum$ denotes summation over index i from one to $h(P)$. Dijkstra's algorithm [21] finds a path with the minimum total cost L . Ranking the retrieved alternatives by their structural characteristics is discussed in Section 3.4.

$$d(u, v) = \frac{1}{\max(w(u, v), \varepsilon)} \quad (6)$$

$$L(P) = \sum_{i=1}^{h(P)} d(v(i-1), v(i)) \quad (7)$$

Figure 1 illustrates the transformation of a document corpus into a contextual semantic network. Its nodes represent terms and validated CVE and ATT&CK identifiers. Undirected links represent concept co-occurrence in documents and retain references to the sources. The resulting network supports weighted search for associative chains.

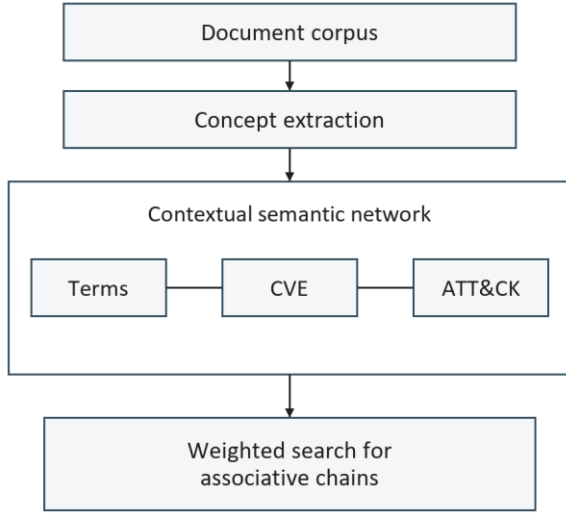


Figure 1: Construction of the contextual semantic network

3.2. Two-Sided Network Construction and Scenario Synthesis

The proposed method is intended to automatically generate alternative scenarios for moving an information system from its current state to a specified target security state. In the general model, a scenario is extracted algorithmically from the semantic network and is then intended to be assessed by a system of virtual experts. This separation defines the structure of synthesis and subsequent explanation of alternatives.

Formally, let S_0 be the current system state, S_T the target state, and $G = (V, E, W)$ the semantic network. A scenario is defined as a directed path $p_i: S_0 \rightarrow S_T$, and the set of possible alternatives is $P = \{p_1, p_2, \dots, p_m\}$.

In this representation, V is the vertex set, E the set of directed links, and W the link attributes. The semantic network defines possible transitions between states, and each scenario corresponds to an admissible path. Applying this formulation to the statistical network in Section 3.1 additionally requires specifying the direction and meaning of the transitions.

A scenario is admissible if all its transitions belong to the semantic network and satisfy the specified constraints R , including technological, temporal, resource, and regulatory constraints. Admissibility checking involves comparing the preconditions of each action with the system state.

Two-sided construction of the contextual network underlies the theoretical formulation.

Forward expansion starts from S_0 and successively includes semantically relevant concepts connected by admissible causal transitions. The process is bounded by search depth, link relevance, and admissibility conditions:

$$G_F = \text{Forward}(G, S_0, R). \quad (8)$$

Backward expansion starts from the target state S_T . It examines concepts and relationships that specify the necessary preconditions for reaching the goal:

$$G_B = \text{Backward}(G, S_T, R). \quad (9)$$

Forward and backward expansion define the search region associated with the initial situation and attainment of the target state. Its size depends on the network structure, expansion depth, and specified constraints.

Subnetwork reconciliation involves identifying shared or semantically equivalent concepts, refining the links between them, and constructing a merged contextual network:

$$G_C = \text{Merge}(G_F, G_B). \quad (10)$$

Reconciliation follows rules for concept equivalence and link compatibility. The network G_C defines the space for subsequent scenario search. Its causal interpretation requires that the original relationships describe the corresponding transitions and retain this meaning after merging.

Alternative scenarios are synthesized by searching for directed paths between S_0 and S_T in network G_C . Cyclic paths, semantically redundant paths, and paths violating the specified constraints are discarded:

$$P = \{p \in \text{Paths}_{G_C}(S_0, S_T) | p \in R\}. \quad (11)$$

In (11), R denotes the set of paths satisfying the specified constraints. The computational procedure may generate a bounded set of alternatives; the completeness of enumeration depends on the algorithm and search conditions. The semantic network determines the scenario structure, while the language model supports subsequent analysis and explanation of the alternatives.

3.3. Multicriteria Assessment

Multicriteria assessment considers a scenario's technical feasibility, risk, implementation time, required resources,

regulatory compliance, and the availability of the required competencies. A set of assessments is formed for each scenario:

$$Q(p_i) = \{q_{i1}, q_{i2}, \dots, q_{ik}\}, \quad (12)$$

The first index of an assessment component denotes the scenario, and the second denotes the criterion; the number of components equals the number of criteria. These assessments determine the aggregate scenario score:

$$J(p_i) = \sum_{j=1}^k \omega_j q_{ij}, \quad (13)$$

where ω_j are nonnegative criterion weights that sum to one. Depending on the task, the weights may be specified by the user or the system configuration. Criterion scores are first normalized to comparable scales with a common direction of preference. The choice of weights and normalization method must be justified for the task. Criterion weights determine each assessment's contribution to the aggregate score and serve a different purpose from semantic network link weights.

Multi-agent analysis provides distributed expert assessment. Specialized agents analyze technical feasibility, risks, resources, regulatory compliance, and possible consequences of a scenario. A corresponding RAG context may be formed for each role. The assessments are reconciled and aggregated; the robustness of the aggregate score requires separate investigation.

The full method is intended to produce a set of alternatives ranked by aggregate score. Each scenario retains its structure, criterion scores, sources, and explanations of the results. This representation enables the analyst to compare the substance of the alternatives.

The general procedure includes updating the semantic network, forward expansion from the current state, and backward expansion from the target state. After subnetwork reconciliation, candidate paths are generated; cyclic, redundant, and inadmissible alternatives are filtered according to the specified constraints. Selected scenarios are passed to specialized agents for assessment. Aggregate scores determine the ranking of scenarios, for which explanations and recommendations are generated.

The method takes as input an infrastructure description, a semantic network, the current context, current and target states, and constraints. The expected output is a ranked set of scenarios

with explanations. Implementing the full cycle requires reconciliation of semantic relationships, action admissibility checks, and integration of graph search with multicriteria assessment.

3.4. Path Search and Ranking in the Prototype

The experimental model implements the alternative-generation stage. Paths are retrieved from the contextual semantic network and compared with reference attack-step sequences. The data, computational settings, and evaluation metrics are presented in Section 4.

The current situation is specified by an infrastructure description and initial concepts, while the goal is the concept representing access obtained at the end of the reference attack. Network nodes represent document concepts, and edges represent their co-occurrence. A retrieved path is interpreted as a hypothesis about a possible relationship between the initial situation and the goal. Converting such a hypothesis into an action sequence requires checking the actions' preconditions.

In the graph model, bidirectional search operates within a single contextual network by expanding regions from the initial nodes and the goal. Their meeting determines the retrieved path. This representation enables investigation of alternative search before implementing the full semantic reconciliation envisaged by the general model.

Correspondence to the reference is assessed using step recall, preservation of step order, and node precision. These metrics characterize the quality of reference-sequence reconstruction. The risk, cost, and regulatory compliance assessments in (12)–(13) require additional input data and experimental validation.

When no initial concepts are supplied, search starts from the goal and covers the reachable network nodes. Candidate initial nodes are selected algorithmically. Retrieved paths are presented in the direction from an initial concept to the goal.

When initial concepts are supplied, a separate bidirectional search is performed for each corresponding node. Each initial node yields at most one path. The results from different initial nodes form a set of alternatives for subsequent ranking.

Alternatives are compared using their mean edge weight. Paths starting at an anchor node are

given priority. Within each group, paths are ordered by decreasing mean weight and, for equal weights, increasing length. This rule defines the structural ranking of alternatives in the contextual network.

The mean weight q is defined by (14). Let $a(P)$ indicate whether the initial node is an anchor and let $s(P)$ denote its normalized name. The key K in (15) is sorted in ascending lexicographic order. The name resolves ties when all other attributes are equal.

$$q(P) = \frac{\sum_{i=1}^{h(P)} w(v(i-1), v(i))}{h(P)} \quad (14)$$

$$K(P) = (1 - a(P), -q(P), h(P), s(P)) \quad (15)$$

Algorithmic alternative generation is separate from language-based explanation. The retrieved paths, structural metrics, and sources provide the context for interpretation by a language model. Their ranking is determined by the graph algorithm. Figure 2 shows the sequence of operations.

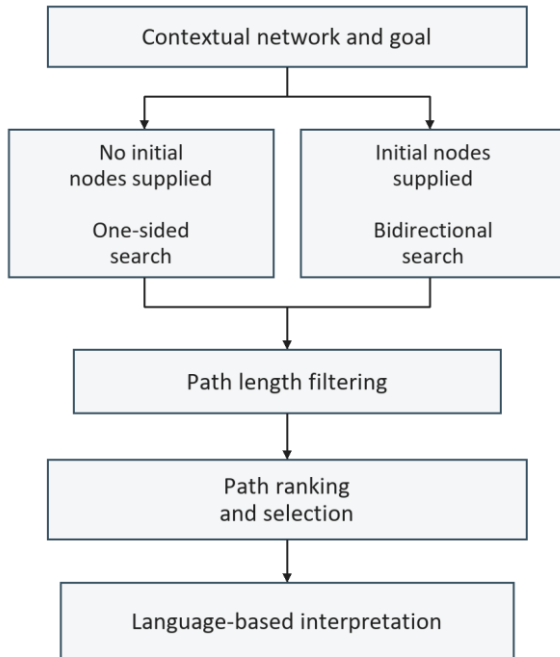


Figure 2: Hypothesis generation with and without supplied initial concepts

Structural ranking characterizes concept connectivity in the contextual network. Multicriteria selection of a defensive scenario requires supplementing this representation with assessments of feasibility, risk, time, and resources, as described in Section 3.3.

3.5. Prototype Architecture and Implementation

The software combines algorithmic data processing with language-based analysis of the results. The algorithmic component constructs the contextual semantic network, computes weights, and searches for paths. The user interface supports selection of the analysis mode and presentation of results. Figure 3 shows the conceptual system architecture. The diagram represents the general model, including causal analysis, multicriteria scenario assessment, and interaction with cyber range infrastructure. The prototype implements the components described in this section; implementation of the full model remains a direction for further development.

3.5.1. Network Preparation and Search Parameters

Text preprocessing precedes construction of the contextual semantic network. Terms are normalized, and stop words and short fragments are discarded. CVE and ATT&CK identifiers are supplied through separate fields and become anchor nodes. Frequencies are based on the number of documents containing a concept or concept pair.

Each document creates links between all concepts included in it. Repeated insertion of a document with the same identifier is skipped, so it contributes to the frequencies only once. This knowledge accumulation procedure supports the addition of new documents; automatically updating an existing description after an infrastructure change requires a separate procedure.

To limit network size, each document contributes up to 40 ordinary terms. More frequent terms are preferred, with a stable selection order for equal frequencies. Anchor nodes are included in addition to this limit. Edge attributes retain up to 200 document references, while the co-occurrence count includes all accepted documents.

In the experimental configuration, the node boost coefficient is two for anchor identifiers and one for ordinary terms. The edge boost is the geometric mean of the coefficients at its endpoints. Edge weight and cost are computed using (4)–(6). Coefficient values are configuration parameters.

NPMI is clipped to the interval from minus one to one. The computation returns minus one when the edge is absent, no documents have been included, or a required frequency is zero. It returns one when the denominator in (3) is no greater than the small positive constant defined in Section 3.1. These conventions handle the boundary cases in which the formula cannot be applied directly or becomes numerically unstable.

The prototype generates alternatives either by one-sided traversal from the goal or by a separate bidirectional search for each recognized initial node. During bidirectional search, it maintains distances from both directions and stops when the optimality condition for the retrieved path is satisfied.

For modes returning several alternatives, minimum-cost paths are found first and then filtered to retain those with one to six hops. The remaining paths are ranked, and up to five are returned. The length limit therefore filters the search results rather than constraining the traversal itself. Section 4 gives the comparison conditions for all three configurations.

3.5.2. Literal Identifier Grounding

Vulnerability and attack technique identifiers are checked during corpus preparation. In the primary mode, an identifier must first occur explicitly in the normalized document text. It is then validated against the National Vulnerability Database (NVD) or the MITRE ATT&CK knowledge base, as appropriate.

A registry match establishes that an identifier exists. Further analysis is needed to determine whether it is relevant to the document and whether the message is reliable. The corpus preparation settings also allow alternative forms of textual grounding.

The accepted identifiers form the document's lists of vulnerabilities and attack techniques. Stored records can be reused without repeating the full processing sequence.

We examined identifier grounding on the system's operational corpus of 1,495 OSINT-channel messages. Before grounding was introduced, the language model had assigned technique identifiers to 379 messages: 380 assignments in total, covering five distinct techniques. T1190 accounted for 375 assignments. None of the identifiers appeared in the text of the message to which it had been

assigned. Registry validation did not reveal this problem because all the identifiers existed in MITRE ATT&CK.

The repeated assignments originated in an example response embedded in the prompt. The example contained specific identifiers, which the model reproduced for unrelated input texts. After these values were removed and textual grounding was introduced, all 380 previous technique assignments failed the check. Pattern matching of identifiers explicitly present in the text increased the number of confirmed vulnerability identifiers from 1 to 75 across 20 documents. This extraction step does not use the language model; the extracted identifiers are subsequently checked against the relevant registries.

We also examined a less restrictive rule: accepting a label when the model provides a quotation found in the document. Two prompts were compared on the same messages. Requiring a quotation eliminated the repeated T1190 label, but the model then assigned other techniques supported by verbatim passages that did not substantiate those assignments. The primary mode therefore requires the identifier itself to appear in the normalized document text.

This evaluation concerns preparation of the OSINT corpus. The path-search experiment in Sections 4.1–4.4 uses 20 test scenarios of cyber intrusions into infrastructure, with identifiers supplied in structured descriptions. It evaluates reference-sequence reconstruction after data preparation and does not test identifier validation.

3.5.3. Result Presentation and Role-Based Analysis

In graph mode, retrieved paths are visualized and passed to the language model for explanation. Search is bidirectional when recognized initial nodes are available and proceeds from the goal otherwise.

The output includes paths, link characteristics, a graph, and a textual explanation. Technique identifiers are displayed with readable names. References attached to links allow the analyst to inspect their sources, and the results can be saved for later analysis.

In role-based mode, active agents receive the same document context in sequence. Each role has its own language model prompt. If several agents return responses, the system produces a combined summary.

Graph search and role-based analysis currently operate as separate components. Integrating them would allow specialized agents to receive retrieved paths for numerical multicriteria assessment.

The contextual semantic network can be saved, reloaded, and exported in GraphML format for reuse or analysis with external tools.

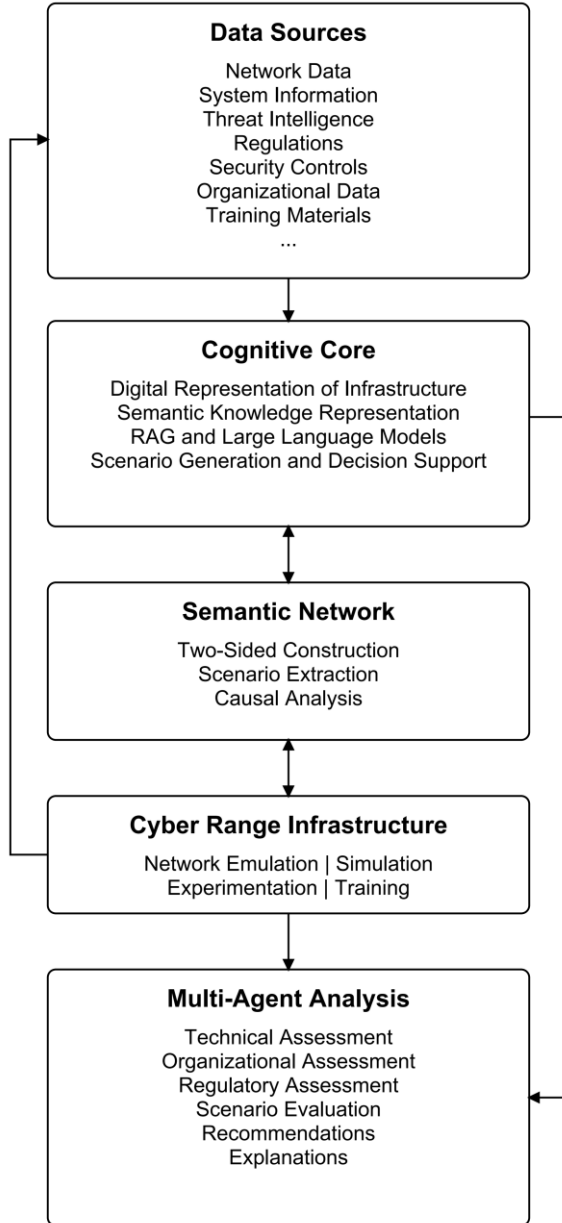


Figure 3: Conceptual system architecture based on a semantic digital twin

4. Experimental Evaluation and Discussion

4.1. Test Scenarios and Data Preparation

The experiment evaluates the correspondence between retrieved paths and reference attack-step

sequences. Three search configurations are compared on 20 test scenarios of cyber intrusions into infrastructure. Each scenario consists of a structured infrastructure description and a reference attack-step sequence. All goal concepts are present in their respective contextual networks.

The test scenarios were prepared from public Hack The Box machine writeups by 0xdf [22]. The dataset is a derived structured representation of these materials. The original experimental result of this study is the comparison of search configurations under the specified evaluation protocol.

Each scenario's infrastructure description is converted into documents about components, identity entities, artifacts, access rights, and vulnerabilities. The reference attack-step sequence is used for evaluation and excluded from the data used to construct the network. The goal is specified by the last meaningful designation of obtained access in the reference sequence. The experiment therefore investigates path search toward a predefined goal.

CVE identifiers are taken from the test scenario descriptions without additional NVD queries. The experimental results therefore depend on the quality of the prepared descriptions; independent validation of the vulnerabilities they contain is outside the protocol.

4.2. Search Configurations and Comparison Conditions

Configuration *A* performs one-sided search from the goal and returns up to five paths. Configuration *B* performs one bidirectional search with a set of initial nodes and returns one path. Configuration *C* performs bidirectional search separately for each initial node, ranks the candidates, and returns up to five paths. Configuration *C* corresponds to the user interface mode with supplied initial nodes.

Initial nodes for *B* and *C* are derived from the names of components whose accessibility field indicates public, VPN, external, anonymous, or unauthenticated access. The first reference step is not used to select initial nodes. Configurations *A* and *C* differ in both the traversal algorithm and the restriction of the initial-node set; their comparison therefore characterizes the configurations as a whole.

All configurations use NPMI-based edge weights. Configurations *A* and *C* return up to five paths with at most six hops. Configuration *B* returns one path without a length limit. Each document contributes up to 40 ordinary terms. Computation involves no language model calls or external network resources.

Repeated computation reproduced the documents, initial nodes, references, paths, and metrics for all 20 test scenarios. Aggregate values agree with the original results. The automated data preparation, path search, and metric computation procedure supports reproducibility under the specified configuration.

4.3. Evaluation Metrics

For evaluation, reference steps are represented as sets of tokens: normalized designations of target entities and obtained access. Entity names are matched to the test scenario description, while action text is excluded from these sets. Empty steps are skipped, and the last nonempty step, which determines the goal, is removed. This representation is used for all subsequent metrics.

Let m be the number of evaluated reference steps, $B(j)$ the token set for step j , and $V(P)$ the node set of the first retrieved path. The set H contains the indices of steps with at least one token match. Reference-step recall C is defined by (16); vertical bars denote set cardinality.

$$H(P) = \{j: V(P) \cap B(j) \neq \emptyset\};$$

$$C(P) = \frac{|H(P)|}{m} \quad (16)$$

For each reference step with a match, r is the first position of a matching token in the retrieved path. The sequence R contains these positions in reference-step order; steps without a match are omitted. The function LNDS gives the length of the longest non-decreasing subsequence. It defines order-preserving recall O in (17). Equal positions are allowed.

$$r(j) = \min\{i: v(i) \in B(j)\};$$

$$O(P) = \frac{\text{LNDS}(R)}{m} \quad (17)$$

Let B denote the union of tokens from all evaluated steps. Node precision Pr in (18) is the fraction of nodes in the first path that belong to this union. All path metrics are zero for an empty result. The evaluated reference sequences are nonempty for all 20 test scenarios.

$$\text{Pr}(P) = \frac{|V(P) \cap B|}{|V(P)|} \quad (18)$$

Let U be the union of nodes from all returned paths. Reference-step recall across up to five paths, C_5 , is defined by (19), where j ranges over the evaluated reference-step indices. This metric does not account for order across different paths. Each test scenario has equal weight in the computation of the mean.

$$C_5 = \frac{|\{j: U \cap B(j) \neq \emptyset\}|}{m} \quad (19)$$

4.4. Reference-Matching Results

For configurations *A* and *C*, the first path in the ranked list and the set of up to five paths are evaluated. The first path has the highest position under the specified ranking rule. The set of up to five paths includes all returned paths when fewer than five are available. Table 1 reports mean values over 20 test scenarios. Recall and precision are computed using (16)–(19).

Table 1

Mean reference-matching metrics

Metric	<i>A</i>	<i>B</i>	<i>C</i>
	One-sided	Single path	Per-start
First path: recall	0.1000	0.1500	0.2417
Order-preserving recall	0.1000	0.1333	0.2250
First path: precision	0.0583	0.1333	0.1625
Up to five paths: recall	0.2042	—	0.6083

Mean reference-step recall across the path set for configuration *C* exceeds that for *A* by 0.4042, or 40.42 percentage points. The ratio of *C* to *A* is approximately 2.98. For the first result, the recall difference is 0.1417. The precision of 0.1625 concerns nodes in the first path for *C*; precision was not separately computed for the union of all paths.

At least one reference step is matched across the retrieved paths in 17 of the 20 test scenarios for *C* and in 6 of 20 for *A*. For the first result, the corresponding counts are 7 and 2; configuration *B* produces matches for 5 scenarios. All reference steps are matched across the path set

for C in seven scenarios, while the first path matches all steps in three. Matches distributed across different paths do not imply reconstruction of the entire sequence by a single path.

Median reference-step recall across the path set for C is 0.5000, with a population standard deviation of 0.3427 over the 20 test scenarios. For A , the median is zero. First-result recall has a median of zero in all three configurations. Considering several alternatives thus increases the number of matched steps, while individual paths still require further validation.

The configurations in Figure 4 are: A , one-sided search from the goal without supplied initial concepts; B , one bidirectional search from a supplied set of initial nodes; and C , separate bidirectional search for each supplied initial node. Panel (a) compares the first paths from all three configurations. Panel (b) compares recall across up to five paths for one-sided search and separate bidirectional searches. The single-search configuration is omitted from panel (b) because it returns only one path. The figure shows mean reference-step recall over 20 test scenarios on a scale from zero to one.

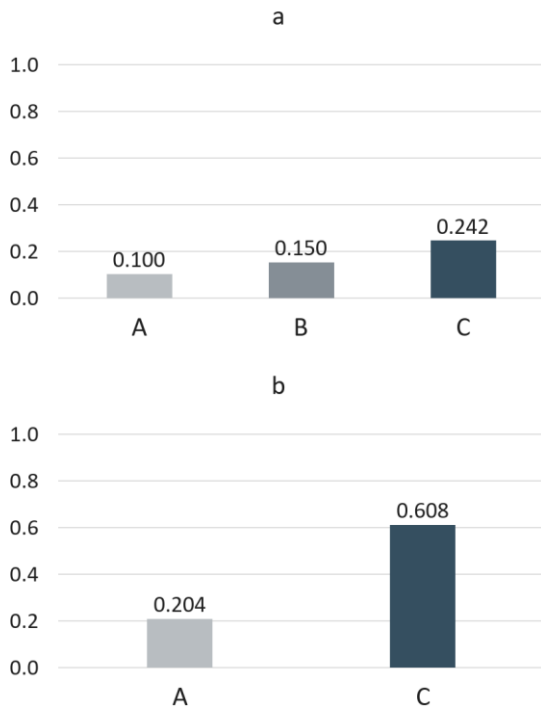


Figure 4: Mean reference-step recall over 20 test scenarios: (a) first path for configurations A, B, and C; (b) up to five paths for configurations A and C.

Table 2

Reference-step recall by reference length

Ref. steps	Scenarios	A	B	C
		Up to five paths	First path	Up to five paths
≥ 1	20	0.2042	0.2417	0.6083
All				
≥ 2	14	0.1488	0.2024	0.5833
Two or more				
≥ 3	4	0.1458	0.3333	0.5417
Three or more				

The dataset contains six one-step references, ten two-step references, three three-step references, and one four-step reference. The rows of Table 2 represent nested subsets: all 20 test scenarios; 14 with at least two steps; and 4 with at least three steps. For longer references, configuration C also has higher recall across the path set. Given the small subset sizes, this finding is descriptive and does not establish a statistically significant advantage on other infrastructures.

4.5. Discussion and Limitations

The scenario-level results reveal differences between the first path and the full set returned by configuration C . In Arctic and Poison, the first path from C has no matches, while the combined path set matches all reference steps. In Forest, first-path recall is 0.6667 and order-preserving recall is 0.3333. Sense, Shocker, and Sunday have no matches even across the full path set from C .

The mean density of the test scenarios' contextual networks is 0.2346. The mean length of the first path from C is 1.55 hops, compared with 1.95 steps in the reference sequence. Concept co-occurrence networks may contain direct links between concepts that are separated by intermediate actions in the reference. This should be considered when interpreting the retrieved paths.

Evaluation relies on token matches. A single match is sufficient to count a reference step as matched, while synonymous designations may

remain unrecognized. The goal token sometimes also occurs in earlier steps, such as administrator in Active and system in Legacy. One node can match several steps, and non-decreasing order allows equal positions. The resulting values therefore characterize correspondence to the adopted reference representation.

The reference sequence is excluded from network construction, although the component descriptions may still contain information about the attack. The experiment thus measures reconstruction of relationships already represented in those descriptions. It does not establish whether the corresponding actions can be executed or whether they improve defense. Those questions require scenario execution in a training environment.

To measure the contribution of bidirectional traversal, both algorithms would need to receive identical initial information. In the present comparison, that information also varies. Reference-matching metrics provide no validation of semantic subnetwork reconciliation or numerical multicriteria assessment; each requires a separate evaluation with suitable data and criteria.

Conclusions

We formulated a semantic digital twin model for cyber defense scenario synthesis and assessment, representing scenarios as admissible paths and defining a general multicriteria assessment procedure. The implemented graph component represents concept co-occurrence in documents and retains references to the documents underlying each link. This component was used for the experimental evaluation.

Across 20 test scenarios of cyber intrusions into infrastructure, mean reference-step recall over up to five paths was 0.6083 for bidirectional search with supplied initial concepts and 0.2042 for one-sided search without them. These values compare the complete configurations, including their different inputs. Reference steps often matched nodes in different alternatives, so recall across the path set could exceed that of the first path. Repeated computation reproduced the reported results.

The prototype allows an analyst to organize documents, retrieve and rank scenario hypotheses, and inspect explanations alongside their sources. Together, these functions support the examination of possible relationships in

cyber threat data. Automatic selection of defensive actions remains a further task. It requires typed causal relationships, checks of action preconditions and effects, experimentally justified assessment criteria, and synchronization with the infrastructure state.

References

- [1] D. Stamatopoulos, M. Katsantonis, P. Fouliras, and I. Mavridis, "Exploring the architectural composition of cyber ranges: A systematic review," *Future Internet*, vol. 16, no. 7, Art. no. 231, 2024.
- [2] J. Zhang, Z. Li, R. Chen, W. Han, N. Hu, and Y. Zhang, "A survey of cyber range: Current status, analysis, and future trends," in *Network Simulation and Evaluation, NSE 2023*. Singapore: Springer, 2024, pp. 88–101, doi: 10.1007/978-981-97-4522-7_6.
- [3] G. Costa, E. Russo, and A. Armando, "Automating the generation of cyber range virtual scenarios with VSDL," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 13, no. 4, pp. 61–80, 2022.
- [4] V. Kampourakis, M. Takaronis, V. Gkioulos, and S. Katsikas, "An ontology-based framework for semantic representation of the cyber range domain," *Journal of Cybersecurity and Privacy*, vol. 6, no. 2, Art. no. 76, 2026.
- [5] E. Cardillo, A. Portaro, and M. Taverniti, "Combining knowledge graph and LLM to extract thesaural relationship and concepts on cybersecurity," in *Proc. ASONAM 2024*. Cham, Switzerland: Springer, 2025, pp. 147–161. [Online]. Available: <https://iris.cnr.it/handle/20.500.14243/513312>
- [6] D. Lande, D. Honcharov, and M. Lytvyn, "A cognitive cyber range based on semantic networks and large language models," SSRN, preprint 7071478, Jul. 2026, doi: 10.2139/ssrn.7071478.
- [7] D. V. Lande, D. K. Honcharov, and M. R. Lytvyn, "A cognitive cyber range based on a semantic digital twin," in *Proceedings of the International Scientific and Practical Conference 'Oxford 2026: Science and Education Today,' Oxford, United Kingdom, Jul. 17–19, 2026*, pp. 72–75 (in Ukrainian).

- [8] D. Lande and L. Strashnoy, "A methodology for system analysis based on semantic digital twins of problem domains," SSRN, preprint 7096618, Jul. 2026, doi: 10.2139/ssrn.7096618.
- [9] D. V. Lande, "Semantic networking for OSINT tasks," in *OSINT Open Source Intelligence: Theory and Methodology*, O. Ye. Korystin and S. V. Demediuk, Eds. Kyiv, Ukraine: 7BC, 2025, pp. 55–57 (in Ukrainian).
- [10] Y. Hong, J. Wu, and R. Morello, "LLM-Twin: mini-giant model-driven beyond 5G digital twin networking framework with semantic secure communication and computation," *Scientific Reports*, vol. 14, Art. no. 19065, 2024.
- [11] T. N. Nguyen, "Toward human digital twins for cybersecurity simulations on the metaverse: Ontological and network science approach," *JMIRx Med*, vol. 3, no. 2, Art. no. e33502, 2022.
- [12] M. A. Ferrag et al., "Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities," *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1–46, 2025.
- [13] S. Ren and S. Chen, "Large language models for cybersecurity intelligence, threat hunting, and decision support," *Computer Life*, vol. 13, no. 3, pp. 39–47, 2025.
- [14] S. Shafee, A. Bessani, and P. M. Ferreira, "Evaluation of LLM-based chatbots for OSINT-based cyber threat awareness," *Expert Systems with Applications*, vol. 261, Art. no. 125509, 2025.
- [15] A. Alhuzali, "LLM-powered threat intelligence: A retrieval-augmented generation approach for cyber attack investigation," *PeerJ Computer Science*, vol. 11, Art. no. e3371, 2025.
- [16] A. Bregar, "Agent-based decision system for proactive cybersecurity strategies," *Journal of Decision Systems*, vol. 35, no. 1, Art. no. 2658199, 2026.
- [17] O. Novikov, D. Lande, and L. Alekseichuk, "Assessment of corporate network penetration scenarios based on iterative link weight determination using the MRRW-PageRank method," in *Proc. Workshop on Cryptology and Data Security (WCDS 2025)*, co-located with SMICS 2025, Lviv, Ukraine, Oct. 16–18, 2025.
- [18] C. R. Landolt, C. Wüsch, R. Meier, A. Mermoud, and J. Jang-Jaccard, "Multi-agent reinforcement learning in cybersecurity: From fundamentals to applications," arXiv:2505.19837, 2025.
- [19] S. Finistrella, S. Mariani, and F. Zambonelli, "Multi-agent reinforcement learning for cybersecurity: Approaches and challenges," in *Proc. WOA 2024, CEUR Workshop Proceedings*, vol. 3735, 2024, pp. 113–128. [Online]. Available: <https://ceur-ws.org/Vol-3735/>
- [20] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *From Form to Meaning: Processing Texts Automatically*, *Proc. Biennial GSCL Conf.*, 2009, pp. 31–40. [Online]. Available: <https://svn.spraakdata.gu.se/repos/gerlof/public/www/Docs/npmi-pfd.pdf>
- [21] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959, doi: 10.1007/BF01386390.
- [22] 0xdf, "Hack The Box machine writeups." [Online]. Available: <https://0xdf.gitlab.io/>