CTF を活用した攻撃成功率の定量化手法の検討

加藤 駿 [1,*]　榎本 聖成 [2]　梶原 聖矢 [2]　原田 竜之介 [2]
藤田 真浩 [1]　鈴木 大輔 [1]　大木 哲史 [2]　西垣 正勝 [2]

概要：製品やサービスのセキュリティ確保において，設計段階でのセキュリティリスクアセスメントの実施が重要視されている．リスクアセスメントでは攻撃データベースを基に膨大な脅威が抽出されるため，費用対効果の観点からガイドラインに基づく優先度付けによる段階的対策が必要である．ガイドラインはリスクアセスメントの知識を形式知化したものであり，脅威抽出や対策優先度の付与を可能とする．しかし，現行のガイドラインでは対策優先度の判断基準が自然言語による曖昧な記述となっている．特に IPA ガイドラインにおいて，対策優先度は資産の重要度や脅威が発生した場合の事業被害，対策の効果，脅威の発生確率から算出されるが，脅威の発生確率の決定変数である攻撃成功率の判断基準が自然言語による曖昧な記述となっていることが課題である．本研究では，セキュリティ教育教材から得られる CTF 問題の正答率等のデータを活用し，攻撃成功率を客観的データから数値化する手法を提案する．本提案により攻撃成功率の客観的数値化が可能となり，リスクアセスメントにおける脅威発生確率を客観的根拠に基づいて導出できるようになる．

キーワード：CTF，リスクアセスメント，攻撃成功率，セキュリティ対策

Investigation of Attack Success Rate Quantification Methods Using CTF

Shun Kato[1,*]　Sena Enomoto[2]　Seiya Kajihara[2]　Ryunosuke Harada[2]
Masahiro Fujita[1]　Daisuke Suzuki[1]　Tetsushi Ohki[2]　Masakatsu Nishigaki[2]

Abstract: In ensuring the security of products and services, the implementation of security risk assessments during the design phase is considered critical. Since risk assessments extract vast numbers of threats based on attack databases, phased countermeasures through prioritization based on guidelines are necessary from a cost-effectiveness perspective. Guidelines represent the formalization of risk assessment knowledge as explicit knowledge, enabling threat extraction and countermeasure prioritization. However, current guidelines present judgment criteria for countermeasure priorities through ambiguous descriptions in natural language. Particularly in the IPA guidelines, countermeasure priority is calculated from asset importance, business damage in case of threat occurrence, countermeasure effectiveness, and threat occurrence probability. Nevertheless, the judgment criteria for attack success rate−which serves as a determining variable for threat occurrence probability−remains problematically described in ambiguous natural language. This study proposes a methodology to quantify attack success rates from objective data by utilizing data such as correct answer rates of CTF problems obtained from security education materials. The proposed approach enables objective quantification of attack success rates and allows threat occurrence probabilities in risk assessments to be derived based on objective evidence.

Keywords: CTF, Risk assessment, Success rate of attacks, Security measures

1. はじめに

セキュリティリスクアセスメント手法は，資産ベースの手法と攻撃シナリオベースの手法に大別される．リスクアセスメントの第 1 フェーズがリスクを含む可能性がある要素の洗い出しである．資産ベースの手法では，システム構成図から資産を洗い出し，攻撃シナリオベースの手法では，システム構成図から保護対象を洗い出す．リスクアセスメントの第 2 フェーズが脅威の抽出である．攻撃を客観的に体系化した攻撃データベースをもとに対象のシステム構成に対してどのような脅威（攻撃）がどの要素に存在するかを抽出する．具体的な攻撃データベースとして，MITRE ATT&CK[1]や CAPEC[2]などが挙げられる．資産ベースの手法では，第 1 フェーズで洗い出した資産に対してどのような脅威(資産被害)が発生する可能性があるかを抽出し，攻撃シナリオベースの手法では，第 1 フェーズで洗い出した保護対象に対してどのような脅威（事業被害）が発生するかを抽出する．リスクアセスメントの結果，膨大な量の脅威が抽出されうる．リスクアセスメントの第 3 フェーズが対策の優先度付与である．通常，すべての脅威に対策を講ずることは費用対効果的に見合わない．ガイドライン[3][4][5]を参考にして，脅威に対して対策すべき優先度を付与し，優先度の高い脅威から順に対策が講じられる．ガイドラインごとに優先度の付与の仕方に若干の違いがある

1 三菱電機株式会社
Mitsubishi Electric Corporation.
2 静岡大学
Shizuoka University.
* kato.shun@bk.mitsubishielectric.co.jp

が，例えばIPA セキュリティリスク分析ガイド[3]では，脅威の大きさ（資産ベースの手法の場合は資産の重要度，攻撃シナリオベースの手法の場合は脅威が発生した場合の被害のレベル），脅威に対する現状の対策の効果，および，脅威が発生する可能性（以下，発生確率）を考慮し，対策優先度を導出する．

すなわち，上記の攻撃データベースやガイドラインは，リスクアセスメントを行うにあたっての知識を形式知化したものであり，その手順に従うことにより脅威の抽出と対策優先度の付与を実施できる．ただし，上記のガイドラインにおいては一般的に，優先度の導出に必要となる脅威の発生確率や対策の効果などの数値の評価基準は，自然言語による定性的な表現に留まっており，リスクアセスメント実施者ごとに異なる値を導出することが考えられる．これらの数値を客観的な情報を用いて定量的に決定することができれば，実施者の能力差に起因するリスクアセスメントの精度変動を抑えることが可能となる．さらには，専門知識を有しない者でも一定水準のリスクアセスメントを実施できるようになる．

本研究では，脅威の発生確率を決定する「攻撃成功率」を，セキュリティ教育教材のデータを用いて定量化することを試みる．具体的には，セキュリティ教育教材の問題正答率を攻撃手法の実行可能性の指標として読み替え，リスクアセスメントの定性的評価を補完するデータ駆動型の評価手法を提案する．

ただし，セキュリティ教育教材の目的はセキュリティスキルやリスクに関する知識の習得であり，現実の攻撃とは観点が異なる．セキュリティ教育教材は具体的な攻撃手法ごとに個別の教育項目として用意されている．一方で，前述（リスクアセスメントの第 2 フェーズ）のとおり，現実の攻撃は攻撃データベースとしてまとめられている．このため，受講者の正答率を攻撃成功率として読み替えるには，セキュリティ教育教材の各教育項目が，攻撃データベースの中のどの攻撃手法に該当する知識であるかを紐づける必要がある．セキュリティ教育教材の教育項目と攻撃データベースの攻撃手法のそれぞれの記載内容等から類似性を見出すことができれば，セキュリティ教育教材の各教育項目と攻撃データベースの各攻撃手法を直接紐づけることが可能である．また，セキュリティ教育教材の教育項目と攻撃データベースの攻撃手法のそれぞれの記載内容等から直接類似性を見出すことができない場合も，媒介変数をもとに両者を間接的に紐づけることが可能である．

本稿では，具体的なセキュリティ教育教材として CTF（Capture The Flag）を活用する．CTF は，各問題において具体的な攻撃手法を実行し，正解となる Flag を見つけることでその問題が正解となる．オープンソースの CTF として picoCTF[6]がある．picoCTF は，セキュリティエキスパートが問題作成を担当しており，（問題によっては）10 万人以上の受講者が存在している．問題ごとの正解者数のみが公開されており，その数値から正答率を求めることができる．具体的な攻撃データベースとして MITRE ATT&CK を活用する．MITRE ATT&CK は攻撃の手法を体系的にまとめたデータベースである．MITRE ATT&CK には攻撃の目的という意味合いの Tactics，各 Tactics を達成するための攻撃手法として Techniques がそれぞれ定義されている．Techniques には，実際の具体的な攻撃手法を示す Procedure（Procedure Examples）と攻撃の緩和策である Mitigation が記載されている．Procedure が実際に行われる攻撃を示すことから，CTF 問題の正答率から Procedure の攻撃成功率を求めることができる．CTF と MITRE ATT&CK を間接的に紐づける場合の媒介変数としては，CWE（Common Weakness Enumeration）[7]を活用する．CWE は，ソフトウェアにおける脆弱性の種類を識別するための共通の基準である．CTF の各問題，MITRE ATT&CK の各 Procedure とそれぞれ関連する CWE 識別子を特定することで，関連する同一の CWE 識別子を介して，CTF の各問題と MITRE ATT&CK の各 Procedure をそれぞれ紐づけることができる．

以降，本稿の構成は以下のとおりである．2 章では，関連動向および既存研究について述べる．3 章では，CTF 問題の正答率を攻撃成功率として読み替え，MITRE ATT&CK の Procedure の攻撃成功率として算出する手法を提案する．4 章では，提案手法の実験および実験結果，考察について述べる．5 章では，本稿のまとめを述べる．

2. 関連動向および既存研究

著者らが調査した限りでは，セキュリティ教育教材の正答率を攻撃成功率の算出に活用した先行事例は見当たらなかった．そこで本章では，リスクアセスメント，攻撃データベース，セキュリティ教育教材に関する関連動向および既存研究を紹介する．

2.1 リスクアセスメント

2.1.1 IPA セキュリティリスク分析ガイド

IPA のセキュリティリスク分析ガイド[3]では，セキュリティのリスクを特定する手順が記載されている．リスクアセスメントには大きく分けて，資産ベースのリスクアセスメント，事業被害（攻撃シナリオ）ベースのリスクアセスメントの二通りである．資産ベースのリスクアセスメントでは，リスクアセスメントを実施する範囲と資産，システム構成，データフローの明確化を行う．資産の重要度，想定する攻撃がどの程度発生するかの指標である脅威レベル，脅威の発生に対する受容可能性の指標である脆弱性レベル，発生した攻撃に対してセキュリティ対策がどの程度有効かの指標である対策レベルをそれぞれ 3 段階で導出し，それらを考慮して，リスク値（対策優先度）を算出する．攻撃シナリオベースのリスクアセスメントでは，リスクアセスメントを実施する範囲と保護対象を明確化する．事業にお

いてどれほどの深刻な被害が出るかの指標である事業被害レベルを定義し，事業被害を引き起こす攻撃シナリオを検討する．攻撃シナリオがどの程度発生するかの指標である脅威レベル，脅威が発生する脆弱性において過去にどの程度被害が発生しているかの指標である脆弱性レベル，発生した攻撃に対してセキュリティ対策がどの程度有効かの指標である対策レベルをそれぞれ 3 段階で導出し，それらを考慮して，リスク値（対策優先度）を算出する．また，脅威レベルは攻撃成功率をもとに評価する方法が記載されている．ただし，脅威レベル，対策レベルはどちらの評価基準も自然言語による定性的な表現に留まっており，定量的な算出手順が明確化されていない．

2.1.2 NIST SP 800-30

NIST SP 800-30 [4]では，資産ベースのリスクアセスメント実施の手引きが記載されている．リスクアセスメントの実施の手引きでは，まず，リスクアセスメントの目的を特定する．次に，懸念される脅威源を特定し，その特徴を定義する．懸念される脅威源から起こりうる脅威を特定し，脅威がもたらす影響およびその可能性を特定する．最後に，起こりうる脅威がもたらす影響，脅威が発生する可能性（発生確率）について考慮され，リスク評価が行われる．ただし，攻撃成功率を用いた脅威の発生確率の具体的な算出方法は明記されていない．

2.1.3 IEC 62443 4-1

IEC 62443 4-1 [5]では，制御システムが満たすべき要件の中に，資産ベースおよび攻撃シナリオベースのリスクアセスメント両方に相当する「脅威モデルを特定する」という手順が含まれている．脅威モデルを作るうえでは，その脅威の影響評価をするような要件も含まれる．その影響を評価する際には，3 段階の定量的評価（低，中，高など）のような単純な方法，可能性と結果に基づいた定量的な方法，CVSS（Common Vulnerability Scoring System）[8]などの標準化された方法の場合があると書かれている．影響を評価するにあたっては，対策度合を定量的に評価する式は定義されているが，攻撃成功率を用いた脅威の発生確率の具体的な算出方法は明記されていない．

2.2 攻撃データベース

攻撃データベースには，MITRE ATT&CK や CAPEC などが存在し，類似のデータベースとしてソフトウェアの脆弱性の種類を識別するための基準である CWE が存在する．本稿においては，提案手法とかかわりの深い MITRE ATT&CK および CWE を以下で詳述する．

2.2.1 MITRE ATT&CK

MITRE ATT&CK[1]は攻撃の手法を体系的にまとめたデータベースである．MITRE ATT&CK には攻撃の目的という意味合いの Tactics, 各 Tactics を達成するための攻撃手法として Techniques がそれぞれ定義されている．Techniques には Procedure（Procedure Examples）と Mitigation が記載されている．Procedure は，実際の攻撃者が使用した具体的な攻撃手法であり，特定の製品における脆弱性と紐づく CVE 番号（CVE-ID）が記載されていることがある．例えば，Technique「Exploit Public-Facing Application」における Procedure は，「APT28 が CVE 2020-0688 や CVE 2020-17144 などの公開されている Exploit を使用して脆弱な Microsoft Exchange を攻撃している．また，外部の Web サイトに対して SQL インジェクション攻撃を行っている．」という脆弱性を使用する旨を示唆する記載があり，Technique「Search Victim-Owned Websites」における Procedure は，「Kimsuky は対象企業のウェブサイトで情報を検索する」という脆弱性を使用しない攻撃手法に関する記載がある．Mitigtion は，Technique を緩和するための緩和策であり，その緩和策の簡易的な説明が記載されている．例えば，Technique「Exploit Public-Facing Application」における Mitigation は，「アプリケーションの分離とサンドボックス化：アプリケーションの分離により，悪用されたターゲットがアクセスできるほかのプロセスやシステム機能が制限される．」という記載がある．

MITRE ATT&CK においては，各 Tactic，各 Technique，各 Procedure に対する攻撃成功率は記載されていない．資産ベースのリスクアセスメントでは，脅威の抽出における粒度は，Tactic, Technique, Procedure のどの粒度でもよいが，粒度が大きいと荒いリスクアセスメントになる一方，抽出される脅威の数が少ないため，対策の選定の作業量は軽くなる．粒度が小さいと詳細なリスクアセスメントになる一方，抽出される脅威が膨大になるため，対策の選定の作業量は重くなることに留意されたい．Tactic，Technique，Procedure をもとにリスクアセスメントを行うためには，各 Tactic，Technique, Procedure の攻撃成功率を算出する必要があるが，この数値化には，専門家であっても経験則による評価となり専門家ごとに異なることが予想され，客観的な数値化は達成できていない．攻撃シナリオベースのリスクアセスメントでは，攻撃シナリオ全体の攻撃成功率を算出する必要がある．Ahmed ら[9]は，攻撃シナリオベースのリスクアセスメントの観点から，MITRE ATT&CK の Tactics および Techniques により構成される攻撃シナリオをアタックツリーで表し，各 Technique に対する攻撃成功率の最大値からクリティカルな攻撃パスを攻撃シナリオ全体の攻撃成功率として算出する方法を提案している．しかし，各 Technique の攻撃成功率の数値化には，上述のとおり，専門家であっても経験則による評価となり専門家ごとに異なることが予想され，客観的な数値化は達成できていない．

2.2.2 CWE（Common Weakness Enumeration）

CWE[7]は，ソフトウェアにおけるセキュリティ上の弱点（脆弱性）の種類を識別するための共通の基準を目指し作成されたものである．CWE では多種多様な脆弱性の種類を脆弱性タイプとして分類し，それぞれに CWE 識別子

（CWE-ID）を付与して階層構造で体系化している．脆弱性タイプは，ある観点から複数の脆弱性タイプを集めた View，共通の特性を持つ脆弱性タイプを集めた Category，個々の脆弱性を表す Weakness，複数の要因が複合した脆弱性を表す Compound Element の 4 種類に分類される．特に Weakness は，最も抽象的な脆弱性の属性である Class，特定のリソースや技術に依存しない脆弱性の属性である Base，個々のリソースや技術，コンテキストなどが特定できるような脆弱性の属性である Variant の 3 つの属性がそれぞれ付与される．各脆弱性タイプには関連する CVE 番号が記載されている．また，NIST NVD（National Vulnerability Database）[10]により，CVE 番号を検索すると，CVE 番号に関する詳細な記載があるページがヒットし，そのページの中には関連する CWE 識別子の記載も存在する．

2.3 セキュリティ教育教材

2.3.1 CTF（Capture The Flag）

CTF はゲーム形式で実施するセキュリティ教育教材であり，セキュリティのスキルを競うコンテストでもある．CTF の各問題には得点がつけられており，各問題で具体的な攻撃を実行することで Flag を見つけられ，その問題で得点することができる．CTF の利用者は最終的な得点をもとに自身のサイバーセキュリティのスキルが世の中の平均と比較してどの程度であるかを把握することができる．ここで，CTF の問題は例えば，ファイルの情報から grep コマンドを用いて Flag を見つけるような脆弱性を使用しない問題から，「SQL インジェクション」の脆弱性を突くことができるかどうかを問う問題，「SQL インジェクションによって窃取した情報を用いての権限昇格」という複数の脆弱性を突くことができるかどうかを問う問題まで様々存在する．また，一般的に，CTF の各問題には解答・解説である Writeup[a]が存在しており，その Writeup の中には，Flag を見つけるために使用する攻撃手法の詳細が記載されている．

オープンソースの CTF として picoCTF[6]がある．picoCTF は，セキュリティエキスパートが問題作成を担当しており，（問題によっては）10 万人以上の受講者が存在している．問題ごとに正解者と受講者による評価（Like）が公開されている．問題のレベルは Easy，Medium，Hard の 3 段階で分けられている．問題のカテゴリは Web Exploitation，Cryptography，Reverse Engineering，Forensics，General Skills，Binary Exploitation の 6 つに分けられている．

3. 提案手法

3.1 概要

2.1 で述べたように，リスクアセスメントにおいて，攻撃成功率を用いた脅威の発生確率（IPA セキュリティ分析ガイドでは「脅威レベル」と呼称）の具体的な算出方法は明記されていない，もしくは，算出方法の記述が自然言語による定性的な表現に留まっており，定量的な算出手順が明確化されていない．本研究では，脅威の発生確率を決定する攻撃成功率を，セキュリティ教育教材のデータを用いて客観的に定量化することを提案する．

一般的な教育教材においては，基礎問題から発展問題まで様々なレベルの問題が用意されており，解答状況から各受講者の知識レベルを把握することができる．セキュリティ教育教材においては，ある具体的な攻撃手法に関する問題に正答した受講者は，その攻撃手法に対する知識を有していると言える．よって，受講者の解答の正誤を攻撃の成否として読み替えることができる．初学者から習熟者まで様々な知識レベルの受講者が存在する．同様に，スクリプトキディからエキスパートまで様々な知識レベルの攻撃者が存在する．セキュリティ教育教材の受講者の知識レベルの分布と攻撃者の知識レベルの分布が一致していると仮定すると，全受講者の正答率を全攻撃者の攻撃成功率として読み替えることができると期待される[b]．

具体的には，2.3.1 で述べたセキュリティ教育教材である CTF 問題の受講者の正答率を用い，2.2.1 で述べた攻撃データベースである MITRE ATT&CK の攻撃成功率として読み替える．ここで，2.3.1 で述べた通り，CTF 問題は具体的な攻撃手法をもとに構成されていることから，「MITRE ATT&CK における具体的な攻撃手法」を示す Procedure と合致する形となる．よって，本研究では CTF の各問題の正答率を MITRE ATT&CK の各 Procedure の攻撃成功率に変換するアプローチを採用する．

図 1 を用いて説明する．紐づけ：CTF 問題が MITRE ATT&CK のどの Procedure に該当する知識であるかを確認して，CTF の各問題と MITRE ATT&CK の各 Procedure とを紐づけて対応関係を整理する．変換：CTF の各問題と MITRE ATT&CK の各 Procedure が紐づけられることによって，各問題の正答率から Procedure の攻撃成功率への変換が可能となる．集約：2.1.1 で述べたように，資産ベースのリスクアセスメントの場合は，Procedure の攻撃成功率を用いて対策優先度を決定することが可能である．一方で，攻撃シナリオベースのリスクアセスメントの場合は，Procedure の攻撃成功率を集約し，Technique の攻撃成功率を導出する．

---

a CTF の各問題に対する Writeup の Web ページは，一般的なインターネット検索により発見可能である．ただし，検索でヒットした Web ページの中から適切な Writeup を発見するには，相応の専門知識を要する．

b 低スキルの攻撃者であっても，既存の攻撃スクリプトやツールを模倣・流用することで攻撃を実行可能性である（いわゆるスクリプトキディ攻撃）．一方，初学者の CTF 参加者も，Writeup を参照して CTF 問題の Flag を得ることができる．これは，攻撃手法を学習・模倣する行為と同等であり，セキュリティ教育教材が「攻撃手法の公開・共有」という現実の脅威環境の一側面をも反映しているという解釈が可能であると考える．

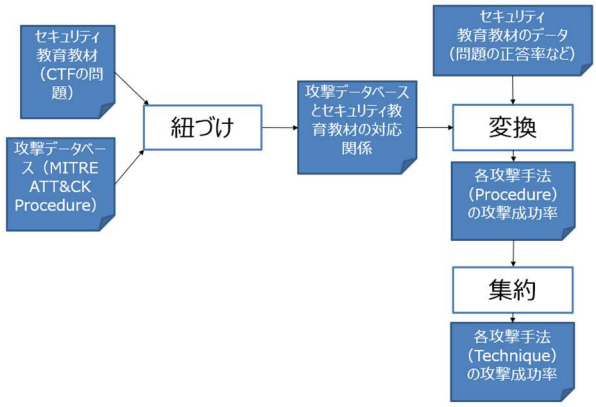


図 1　攻撃成功率導出の全体像

3.2 CTF と MITRE ATT&CK の紐づけ

3.2.1 Procedure の分類

Procedure は，まず，Exploit 型/非 Exploit 型で分類することができる．Exploit 型 Procedure は，脆弱性を使用する攻撃手法である．非 Exploit 型 Procedure は，Exploit の前準備（偵察，調査など）や後段階（機密情報の窃取など）の脆弱性を使用しない攻撃手法である（エリア①）．Procedure は実際に存在する攻撃団体などが行った攻撃事例をもとに記載されるため，未知の脆弱性の分類となる Procedure は存在しない．既知の脆弱性には，CVE 番号と関連する CWE 識別子が付与される．よって，Exploit 型の Procedure はさらに，CVE 番号の記載あり/なし，CWE 識別子の関連あり/なしcでそれぞれ分類することができる．すなわち，「CVE 番号の記載あり+CWE 識別子の関連あり」の Procedure（エリア②），「CVE 番号の記載なし+CWE 識別子の関連あり」の Procedure（エリア③），「CVE 番号の記載あり+CWE 識別子の関連なし」の Procedure（エリア④），「CVE 番号の記載なし+CWE 識別子の関連なし」の Procedure（エリア⑤）で分類することが可能である（図 2）．

【エリア①】

非 Exploit 型 Procedure は，脆弱性特定ラベルによる体系化がなされていないため，個々の攻撃手法の説明がそれぞれ具体的な記述レベルで，Procedure の中に自然言語で記載される形となる．また，CTF 問題の Writeup は，一般的に問題の解答だけでなくその解説まで示されているため，個々の CTF 問題の説明がある程度詳細な記述レベルで，Writeup の中に自然言語で記載される形となる．よって，自然言語処理に優れた LLM を用いて，両者の紐づけを適切に推論できると期待される．この LLM を Q2P_LLM と呼称する．この LLM を用いて，Procedure と CTF 問題の直接紐づけが可能となる．

【エリア②】

「CVE 番号の記載なし+CWE 識別子の関連あり」の Exploit 型 Procedure は，脆弱性特定ラベルによる体系化がなされており，脆弱性名などが Procedure の中に自然言語で記載されている．また，JVN iPedia（脆弱性対策情報データベース）や CWE のサイトにおいて，個々の CWE 識別子の説明がそれぞれ自然言語で記載されている．さらに，CTF 問題の Writeup は，一般的に問題の解答だけでなくその解説まで示されているため，個々の CTF 問題で用意されている脆弱性の説明がある程度詳細な記述レベルで，Writeup の中に自然言語で記載される形となる．よって，自然言語処理に優れた LLM を用いて，Procedure と関連する CWE 識別子，ならびに，CTF 問題と関連する CWE 識別子をそれぞれ適切に特定できると期待される．前者の LLM を P2C_LLM，後者の LLM を Q2C_LLM と呼称する．これら 2 つの LLM を用いて，CWE 識別子を介した Procedure と CTF 問題の間接紐づけが可能となる．

【エリア③】

「CVE 番号の記載あり+CWE 識別子の関連あり」の Exploit 型 Procedure は，NIST NVD により CVE 番号から演繹的に CWE 識別子を特定可能である．現在の LLM のハルシネーションに鑑み，本研究では，CWE 識別子を演繹的に特定可能である Procedure においては，演繹的に特定することとする．この演繹識別器を P2C_Linker と呼称する．よって，P2C_Linker とエリア②の Q2C_LLM を用いて，CWE 識別子を介した Procedure と CTF 問題の間接紐づけが可能となる．

【エリア④】

「CVE 番号の記載あり+CWE 識別子の関連なし」の Exploit 型 Procedure は，脆弱性の詳細な情報が公開されていない，もしくは，わかっていない Procedure である．CTF 問題は具体的に判明している脆弱性などをもとに問題を作成するため，当該 Procedure と紐づく CTF 問題は存在しないこととなる．エリア④の Procedure については，今後，他のセキュリティ教育教材との紐づけの可能性を検討していく．

【エリア⑤】

「CVE 番号の記載なし+CWE 識別子の関連なし」の Exploit 型 Procedure は，付番されていない CWE 識別子と関連する Procedure であり，例えば，「Andariel は、ゼロデイ脆弱性を含む多数の ActiveX の脆弱性を悪用した」といった内容の記載に留まる．一般的に，特定の製品（上の例では ActiveX）に対する CTF 問題は存在しない．よって，エリア④と同様，今後の課題とする．

c CWE 識別子の関連ありとは，付番されている CWE 識別子が関連している場合であり，CWE 識別子の関連なしとは，CWE-noinfo などの付番されていない CWE 識別子が関連している場合である．

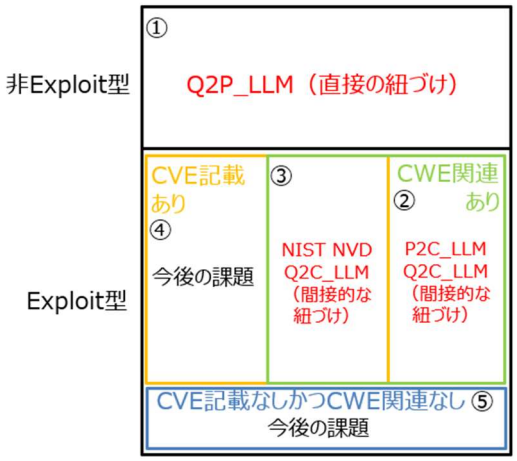


図 2　Procedure の分類

3.2.2 エリア①における CTF と MITRE ATT&CK の紐づけ

Q2P_LLM を用いたエリア①の紐づけにおいては，「CTF 問題 → Procedure」，「CTF 問題 ← Procedure」の二つのアプローチがある．LLM はインターネットの情報を学習に利用しているため，インターネット上で広く情報が存在する物事に関するほど，より正確な知識回答装置として機能すると期待される．MITRE ATT&CK は攻撃を体系化したデータベースであるため，インターネット上の様々な記事やサイトで引用されている．よって，本手法では CTF 問題を LLM に与え，その問題と紐づく MITRE ATT&CK の Procedure を得るアプローチを採用する（図 3 の①）．Q2P_LLM のインターフェースは以下のとおりである．

【Q2P_LLM】

入力：CTF 問題の Writeup 記載内容

出力：当該問題と関連する Procedure（複数可）

LLM の代表的なプロンプトエンジニアリング手法には，RAG（Retrieval-Augmented Generation），ICL（In-Context Learning），CoT（Chain of Thought）がある[11]．したがって，Q2P_LLM は，これら 3 種類の手法を組み合わせて構成する．ここで，Procedure に関する情報は，主に MITRE ATT&CK のサイトに存在するため，RAG のインデックスには当該サイトを指定する．

3.2.3 エリア②における CTF と MITRE ATT&CK の紐づけ

Q2C_LLM および P2C_LLM を用いたエリア②の紐づけにおいては，「CTF 問題 → CWE 識別子 → Procedure」，「CTF 問題 ← CWE 識別子 ← Procedure」，「CTF 問題 → CWE 識別子 ← Procedure」の三つのアプローチがある．本手法では，三つ目のアプローチを採用する（図 3 の②）．理由としては，異なる CTF 問題や Procedure が同じ脆弱性を指し示す場合があるため，その際には「CWE 識別子 → Procedure」あるいは「CTF 問題 ← CWE 識別子」の特定が一意に定まらないことが挙げられる．　Q2C_LLM，P2C_LLM それぞれのインターフェースは以下のとおりである．

【Q2C_LLM】

入力：CTF 問題の Writeup 記載内容

出力：当該問題と関連する CWE 識別子（複数可）

【P2C_LLM】

入力：Procedure の記載内容

出力：当該 Procedure と関連する CWE 識別子（複数可）

Q2C_LLM，P2C_LLM はどちらも，Q2P_LLM と同様，RAG，ICL，CoT の組み合わせにより構成する．ただし，RAG のインデックスには JVN iPedia や CWE サイトを指定する．二つの LLM により同一の CWE 識別子が特定される場合，CWE 識別子を介して CTF 問題と Procedure が紐づけられる．

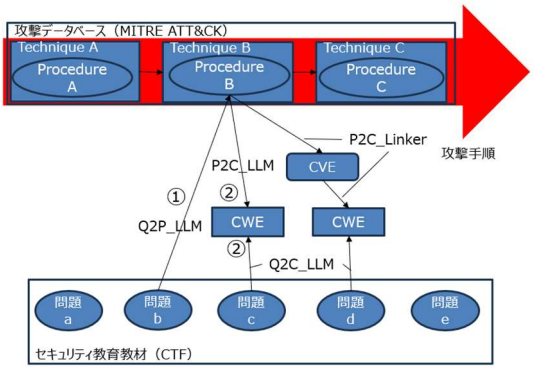
3.2.4 エリア③における CTF と MITRE ATT&CK の紐づけ

エリア③の紐づけにおいては，エリア②における P2C_LLM が P2C_Linker に置き換わる．P2C_Linker は，Procedure から CWE 識別子を演繹的に特定するツールである．具体的には，Procedure に記載された CVE 番号を NIST NVD で検索し，その詳細ページから CWE 識別子を取得する．P2C_Linker のインターフェースは以下のとおりである．

【P2C_Linker】

入力：Procedure に記載されている CVE 番号

出力：当該 Procedure と関連する CWE 識別子（複数可）



3.3 図 3　本手法における紐づけのアプローチ CTF の正答率から攻撃成功率への変換

3.3.1 Procedure の攻撃成功率の計算方法

3.2 の手順により，CTF 問題と MITRE ATT&CK Procedure が紐づけられた．しかし，CTF 問題と MITRE ATT&CK の Procedure の関係は多対多となりうる（図 4）ため，CTF 問題の正答率から MITRE ATT&CK の Procedure の攻撃成功率への変換が必要である．その方法は三通りに大別される（図 5）．一つ目は，一対一で紐づいている場合において，CTF 問題の正答率をそのまま MITRE ATT&CK の Procedure の攻撃成功率[d]とする方法（方法 P1）である．二つ目は，多対一で紐づいている場合[e]において，複数の CTF 問題の正

d 攻撃成功率を攻撃難易度として表現したい場合には，今回は単純に「攻撃難易度 ＝1－ 攻撃成功率」という変換を用いることとする．

e 一般的に教育教材においては，1 つの単元に対して複数の問題が用意されるため，1 つの Procedure に複数の CTF 問題が紐づくことになる．

答率を何らかの方法で計算して MITRE ATT&CK の Procedure の攻撃成功率へと変換する方法（方法 P2）である．三つ目は，一対多で紐づいている場合fにおいて，CTF 問題の正答率を何らかの方法で計算して，紐づいている MITRE ATT&CK の Procedure それぞれの攻撃成功率へと変換する方法（方法 P3）である．

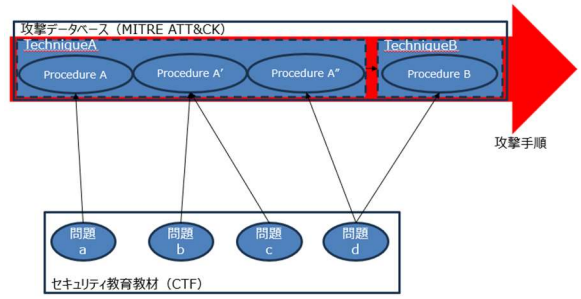


図 4　変換の全体像

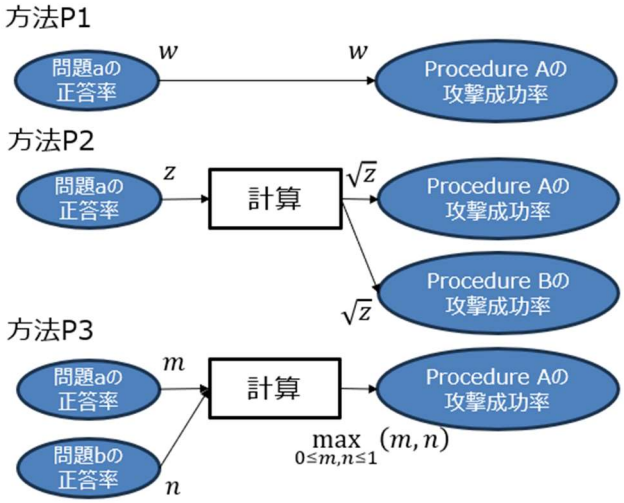


図 5　Procedure の攻撃成功率に変換する方法

方法 P1 は，方法・手順ともに自明である．方法 P2 では，CTF 問題の正答率を複数の Procedure の攻撃成功率に計算する必要がある．例えば，CTF 問題 a に二つの異なる種類の Procedure が紐づいていた場合，Procedure A の攻撃成功率をx，Procedure B の攻撃成功率をy，CTF 問題 a の正答率をzと置くと，$z = x * y$の式となり，x，yを求める計算はzの因数分解を求めればよい．ただし，因数分解は何通りもの解をもつ場合があるため，正確な値を導出することは困難である．本稿では，因数分解の解を求めるために，紐づいている複数の Procedure の攻撃成功率が等価であると仮定し，$x = y(= \sqrt{z})$として計算する．CTF 問題に二つの同じ種類の Procedure が紐づいていた場合，CTF 問題の正答率をそのまま二つの Procedure それぞれの攻撃成功率とする．方法 P3 では，紐づいている複数の CTF 問題の正答率から各 Procedure の攻撃成功率を算出する手段が必要である．今回は，Weakest Link の考え方に基づき，Procedure に紐づいている複数の CTF 問題の正答率の最大値を当該 Procedure の攻撃成功率とする．

3.3.2 Technique の攻撃成功率への集約

3.1 で述べたように，資産ベースのリスクアセスメントでは，Procedure の攻撃成功率を用いて対策優先度を決定できる．一方，攻撃シナリオベースのリスクアセスメントでは，対策優先度の決定に Technique の攻撃成功率を用いるため，Procedure の攻撃成功率を Technique の攻撃成功率に集約する必要がある．Procedure と Technique の関係は多対一となりうる（図 4 の点線内）ため，その方法は二通りに大別される．一つ目は，一対一となる（Technique を実現する Procedure が 1 つのみ存在する）場合において，Procedure の攻撃成功率をそのまま Technique の攻撃成功率とする方法（方法 T1）である．二つ目は，多対一となる（Technique を実現する Procedure が複数存在する）場合において，複数の Procedure の攻撃成功率を何らかの方法で計算して Technique の攻撃成功率へと変換する方法（方法 T2）である．

方法 T1 は，方法・手順ともに自明である．方法 T2 では，Technique を実現する複数の Procedure の攻撃成功率を集約し，その Technique の攻撃成功率を算出する手段が必要である．今回は，Procedure の攻撃成功率の計算方法と同様に Weakest Link の考え方に基づき，Technique を実現するすべての Procedure の攻撃成功率の最大値を当該 Technique の攻撃成功率とする．

4. 実験・実装

4.1 実験概要

基礎実験を通じて，3 で説明した提案手法（CTF 問題の正答率から MITRE ATT&CK Procedure の攻撃成功率の算出）の実現可能性を検証する．ここで，CTF 問題と Procedure の紐づけについては，3.2.1 で分類したエリア②に該当する Procedure を実験対象とした．エリア②は「CVE 番号の記載があり，CWE 識別子が関連付けられている Procedure」であり，他のエリアの Procedure と比較して多くの情報がインターネット上に存在するため，提案手法の基本的な有効性を確認する第一段階として適していると判断した．

4.2 実験システムの仕様

実験システムの仕様を以下に示す．

- **目的**：CTF 問題の正答率を用いた Procedure の攻撃成功率算出システムの構築
- **対象 Procedure**：MITRE ATT&CK Enterprise に分類される全 Procedure のうち，CVE 番号が記載され，

f 例えば「SQL インジェクションによって窃取した情報を用いての権限昇格」を問う CTF 問題には，SQL インジェクションと権限昇格の 2 つの異なる種類の Procedure が紐づくことになる．また，「SQL インジェクション」の攻撃ができるかどうかを問う CTF 問題に，「攻撃者 XX が SQL インジェクションを実行」や「攻撃者 YY が SQL インジェクションを実行」など複数の同じ種類の Procedure が紐づくこともある．

かつ CWE 識別子と関連性を有するもの

- **対象 CTF 問題**：picoCTF の問題群から，Forensics および Reverse Engineering カテゴリを除外し[g]，以下のリポジトリに Writeup が記載されている問題 https://github.com/Cajac/picoCTF-Writeups/tree/main/

4.3 実験システムの実装

実験システムは以下の手順で Procedure の攻撃成功率を算出する．

(1) **Procedure と関連する CWE 識別子の特定：P2C_Linker**

MITRE ATT&CK が提供する STIX データを用いて，CVE 番号が記載された Procedure を抽出する．続いて，NIST NVD API を利用して抽出された CVE 番号と関連する CWE 識別子を特定し出力する．

(2) **CTF 問題と関連する CWE 識別子の特定：Q2C_LLM**

CTF 問題の Writeup から関連する CWE 識別子を特定する RAG，ICL，CoT をそれぞれ構築する．3 種類の LLM が出力する CWE 識別子と Writeup の内容をもとに，第 4 の LLM で各 CWE 識別子の確信度を算出する．その確信度に閾値を設定し，閾値以上の確信度を有する CWE 識別子を各 CTF 問題の関連 CWE 識別子として出力する．

(3) **CTF 問題と Procedure の紐づけ**

（1）および（2）で得られた Procedure および CTF 問題それぞれと関連する CWE 識別子を突合し，同一の CWE 識別子を有する Procedure と CTF 問題の対応関係を出力する．

(4) **Procedure の攻撃成功率の算出**

3.3.1 で説明した方法 P1，P2，P3（図 5）をこの手順で実行することにより，各 Procedure に紐づけられた CTF 問題の正答率から当該 Procedure の攻撃成功率を算出する．

4.4 倫理的配慮

本実験システムでは，CWE サイトで配布されている XML ファイル，MITRE ATT&CK の STIX データ，および NIST NVD の API を利用する．取得対象は公開されている技術情報のみに限定している．NIST NVD へのアクセスにおいては，API 利用制限の遵守および適切なアクセス間隔の設定により，過度な負荷を回避している．CTF の Writeup および問題の正答率については，手動でデータ収集を実施することにより，サイトへの不適切なアクセスを防止し，利用規約の遵守を確保している．

4.5 実験結果

- 対象 Procedure 数：134 件
- 対象 CTF 問題数：162 件
- 紐づけが確認された Procedure 数：91 件
- 紐づけ確認率：67.9%
- 算出された攻撃成功率の範囲：0.5～65.15%

5. おわりに

本稿では，セキュリティ教育教材と攻撃データベースの紐づけ手法，および教育教材を活用した攻撃成功率（攻撃難易度）の客観的数値化手法を提案した．従来のリスクアセスメントにおいては，攻撃成功率の算出手順が自然言語による定性的記述に留まり，定量的算出プロセスが明確化されていなかった．その結果，セキュリティ専門家であっても客観的かつ再現可能な攻撃成功率評価の実施が困難であるという課題が存在していた．本研究の提案手法により，この課題の解決に向けた基盤技術を提供することができた．

今後の課題は以下のとおりである．（イ）4.5 に示した実験結果の妥当性の評価．（ロ）エリア③以外の Procedure の攻撃成功率の算出．（ハ）アルゴリズムの改良（Weakest Link 以外の攻撃成功率算出，多対多の紐づけに対する攻撃成功率の調整，CTF 問題の受講者数の推定）．（ニ）利用するセキュリティ教育教材の拡充（様々な Writeup の使用，picoCTF 以外の CTF の利用，CTF 以外のセキュリティ教育教材の活用）．

参考文献

[1] MITRE ATT&CK, https://attack.mitre.org/, (参照 2024-07-17).
[2] CAPEC – Common Attack Pattern Enumeration and Classification, https://capec.mitre.org/, (参照 2024-07-17).
[3] IPA 独立行政法人 情報処理推進機構, 制御システムのセキュリティリスク分析ガイド第２版, https://www.ipa.go.jp/security/controlsystem/ssf7ph00000098vy-att/000109380.pdf, 2023/03.
[4] NIST SP 800-30（IPA 翻訳），リスクアセスメント実施の手引き, https://www.ipa.go.jp/security/reports/oversea/nist/ug65p90000019cp4-att/000025325.pdf, 2012/09.
[5] IEC 62443 4-1, https://webstore.iec.ch/publication/33615, 2018.
[6] picoCTF, https://www.picoctf.org/, (参照 2024-07-17).
[7] IPA 共通脆弱性タイプ一覧 CWE 概説, https://www.ipa.go.jp/security/vuln/scap/cwe.html, (参照 2024-03-13).
[8] 共通脆弱性評価システム CVSS v3 概説, https://www.ipa.go.jp/security/vuln/scap/cvssv3.html, (参照 2024-08-07).
[9] M. G. Ahmed, S. Panda, C. Xenakis and E. Panaousis, MITRE ATT&CK-driven Cyber Risk Assessment, Proceedings of the 17th International Conference on Availability, Reliability and Security (ARES'22), no.107, pp.1-10, Aug. 2022.
[10] NATIONAL VULNERABILITY DATABASE, https://nvd.nist.gov/, (参照 2024-08-13).
[11] M. Son, S. Lee, Performance Analysis of Prompt-Engineering Techniques for Large Language Model, 2025 IEEE International Conference on Consumer Electronics (ICCE),

g Forensics の CTF 問題を除く理由として，Forensics の CTF 問題は与えられたログファイルなどから攻撃の痕跡を探すパターンが主であるため，脆弱性と直接のかかわりがないこと，Reverse Engineering の CTF 問題を除く理由として，Reverse Engineering の CTF 問題は与えられたプログラムファイルを解析するパターンが主であるため，脆弱性と直接のかかわりがないことが挙げられる．