POLITECHNIKA WARSZAWSKA

DYSCYPLINA NAUKOWA INFORMATYKA TECHNICZNA
I TELEKOMUNIKACJA
DZIEDZINA NAUK INŻYNIERYJNO TECHNICZNYCH

Rozprawa doktorska

mgr inż. Marek Bogusław Janiszewski

Metodyka oceny wiarygodności systemów zarządzania zaufaniem i reputacją

Promotor
dr hab. inż. Krzysztof Szczypiorski, prof. uczelni

WARSZAWA 2023

Składam serdeczne podziękowania Panu dr. hab. inż. Krzysztofowi Szczypiorskiemu, prof. PW za opiekę merytoryczną, wsparcie i cenne sugestie. Jestem przekonany, że zdobyte doświadczenie i umiejętności przyczynią się do mojego dalszego rozwoju zawodowego.

Dziękuję Panu prof. dr. hab. inż. Józefowi Lubaczowi za opiekę merytoryczną w początkowej fazie badań.

Dziękuję również moim kolegom z NASK-PIB, na czele z dr. inż. Adamem Kozakiewiczem i dr inż. Anną Felkner za pierwsze recenzje fragmentów rozprawy i cenne uwagi oraz dyrektor ds. naukowych NASK-PIB, Pani prof. dr hab. inż. Ewie Niewiadomskiej-Szynkiewicz za wsparcie i motywację, a także Panu prof. dr. hab. inż. Markowi Amanowiczowi za owocne dyskusje.

Szczególne podziękowania składam moim Rodzicom, za wykształcenie we mnie ciekawości świata i chęci ciągłego rozwoju oraz za stworzenie warunków do ich realizacji na początkowych etapach mojego życia. Dziękuję także mojej Wybrance serca oraz całej Rodzinie i Bliskim za wsparcie i wyrozumiałość.

METODYKA OCENY WIARYGODNOŚCI SYSTEMÓW ZARZĄDZANIA ZAUFANIEM I REPUTACJĄ

STRESZCZENIE

Systemy zarządzania zaufaniem i reputacją, jako mechanizmy miękkiego bezpieczeństwa, znajdują zastosowanie w wielu środowiskach teleinformatycznych. Systemy te, mimo że stanowią zabezpieczenie przeciwko atakom związanym ze świadczeniem nierzetelnych usług, to same mogą stać się celem specyficznych dla nich ataków. W związku z tym istnieje potrzeba kompleksowej oceny wiarygodności (rozumianej jako odporności na ataki) takich systemów. W niniejszej rozprawie przedstawiono ogólną ideę systemów zarządzania zaufaniem i reputacją, pojęcia z nimi związane, a także przykłady ich praktycznych zastosowań. Zawarto także przegląd stanu wiedzy w zakresie ataków na systemy zarządzania zaufaniem i reputacją oraz oceną ich wiarygodności, w tym odniesienie się do publikacji, wchodzących w polemikę z wcześniejszymi artykułami autora rozprawy, dotyczącymi tej tematyki. W pracy zaprezentowano metodykę oceny wiarygodności, bazującą na stworzonych od podstaw i przedstawionych w rozprawie modelach środowiska, systemów zarządzania zaufaniem i reputacją oraz ataków. W metodyce zdefiniowano miary wiarygodności, które mogą służyć ocenie odporności systemów na ataki, a także przedstawiono rodzaje i propozycje badań, które mogą zostać wykorzystane do ich ewaluacji. Zaprezentowano także propozycję metody pozwalającej na identyfikację nowego ataku przeciwko konkretnemu systemowi zarządzania zaufaniem i reputacją, który może być bardziej efektywny niż znane do tej pory ataki. Rozprawa zawiera także opis dedykowanego narzędzia stworzonego do ewaluacji wiarygodności systemów zarządzania zaufaniem i reputacją. W oparciu o stworzoną metodykę i narzędzie, zaprezentowano wyniki przeprowadzonych badań wybranego systemu oraz ocenę jego wiarygodności, co stanowi przykład praktycznego zastosowania modeli, metodyki i metody przedstawionych w niniejszej publikacji. W ramach podsumowania zostały zaprezentowane perspektywy kontynuacji badań, a także wskazane dotychczasowe publikacje autora dotyczące ataków na systemy zarządzania zaufaniem i reputacją oraz oceny ich wiarygodności.

Słowa kluczowe: *zaufanie, reputacja, ataki, systemy zarządzania zaufaniem i reputacją, systemy zaufania, systemy reputacyjne, mechanizmy miękkiego bezpieczeństwa*

RELIABILITY EVALUATION METHODOLOGY OF TRUST AND REPUTATION MANAGEMENT SYSTEMS

ABSTRACT

Trust and reputation management systems, as a soft security mechanism, are used in many information and telecommunication environments. These systems, although they can provide protection against attacks usually associated with the provision of unreliable services, can themselves become the target of attacks specific to them. Therefore, there is a need for a comprehensive evaluation of the reliability (the resistance to attacks) of such systems. This dissertation presents the general idea of such systems, concepts related to them, as well as examples of their practical applications and attacks on this type of systems. It also contains an overview of the state of the art in the field of attacks on trust and reputation management systems and the assessment of their reliability, including references to publications that enter into polemics with earlier articles created by the author of the dissertation regarding this subject. On the basis of the formal models of the environment, the trust and reputation management system, and the attack on such a system, the dissertation presents a methodology for assessing their reliability. The methodology includes the definitions of reliability measures that can be used to evaluate the resistance of systems to attacks, and also presents the types and proposals of research that can be used to evaluate such systems. The work also includes a proposal of a method that allows the identification of new attacks against the specific trust and reputation management system, which may potentially be more effective than well-known attacks. The dissertation also contains a description of a dedicated testbed created to evaluate the reliability of trust and reputation management systems. Based on the created methodology and testbed, the work presents the results of the tests of a system and the assessment of its reliability, which is an example of the practical application of the models, methodology and method presented in this publication. As part of the summary, the prospects for the continuation of research are presented, as well as the author's previous publications regarding the attacks on trust and reputation management systems and the assessment of their reliability.

Keywords: *trust, reputation, attacks, trust and reputation management systems, trust systems, reputation systems, soft security*

SPIS TREŚCI

1. WPROWADZENIE

1.1. MOTYWACJA I KONTEKST PRACY

Mechanizmy zapewnienia bezpieczeństwa systemów, sieci i usług teleinformatycznych stanowią przedmiot zainteresowania wielu badań, projektów i przedsięwzięć. Dynamiczny postęp w tej dziedzinie napędzany jest z jednej strony ciągłym zwiększeniem uzależnienia współczesnego społeczeństwa od teleinformatyki i jej znaczenia w codziennym życiu, a z drugiej strony ewolucją zagrożeń. Oprócz różnorakich sposobów realizacji standardowych usług bezpieczeństwa, tj. poufności, integralności i dostępności, stosowane są inne metody mające zapewnić rzetelność działań użytkowników systemu lub sprawiedliwość podziału zasobów w określonym systemie, sieci lub usłudze.

Środowiska teleinformatyczne są złożone z wielu agentów świadczących usługi na rzecz innych agentów. Ewolucja usług teleinformatycznych wymusza zmianę sposobów zapewnienia bezpieczeństwa, ze względu na to, że stosowane zabezpieczenia muszą chronić przed nowymi atakami i zagrożeniami. Celem zarządzania zaufaniem i reputacją (TRM – ang. „Trust and Reputation Management") jest zapewnienie większej efektywności, pomimo istnienia agentów dążących do maksymalizacji własnej użyteczności, a nie użyteczności środowiska jako całości. Zastosowania zarządzania zaufaniem i reputacją obejmują różnorodne środowiska, takie jak systemy rekomendacyjne, serwisy aukcyjne, aplikacje mobilne, systemy detekcji i filtracji spamu, systemy monitorowania reputacji adresów IP w celu wykrywania różnych ataków w Internecie, w tym ataków odmowy usług, czy systemy oceny rzetelności informacji. Systemy TRM są też wykorzystywane do poprawy efektywności działania protokołów routingu, sieci sensorowych i mobilnych, a także innych typów sieci.

Systemy zarządzania zaufaniem i reputacją należą do klasy tzw. mechanizmów miękkiego bezpieczeństwa, które opierają się na subiektywnych ocenach uczestników systemu, w odróżnieniu od mechanizmów tzw. twardego bezpieczeństwa, opartych głównie na środkach kryptograficznych i kontroli dostępu. Celem stosowania mechanizmów, zarówno miękkiego jak i twardego bezpieczeństwa, jest obrona przed działaniami nierzetelnych agentów. W przypadku mechanizmów miękkiego bezpieczeństwa, obrona polega na ograniczaniu liczby interakcji z nierzetelnymi agentami (np. żądań wykonania pewnej usługi), podczas gdy mechanizmy twardego bezpieczeństwa mają za zadanie zapobieganie nawiązaniu jakiejkolwiek komunikacji przez agentów niemających odpowiednich praw dostępu.

Pomimo szerokiego stosowania w wielu typach środowisk, sieci i systemów teleinformatycznych, zagadnienia dotyczące systemów zarządzania zaufaniem i reputacją

znacząco wykraczają poza podstawowy przedmiot zainteresowania teleinformatyki. Świadczą o tym nie tylko zastosowania tego typu systemów, ale również ich geneza. Pojęcia zaufania i reputacji, będące podstawą idei takich systemów, wyewoluowały z nauk społecznych, przede wszystkim z socjologii i ekonomii, ale także socjocybernetyki, czyli subdyscypliny z pogranicza cybernetyki i teorii systemów zajmującej się analizą zachowań jednostek jako elementów tworzących systemy społeczne. W analogii do członków grupy społecznej nawiązujących relacje, środowisko składa się z agentów podejmujących interakcje i świadczących określone usługi (np. udostępnianie plików, przekazywanie pakietów, dostarczanie odczytów z sensorów). Złożone relacje społeczne częściowo charakteryzuje zaufanie pomiędzy parą członków grupy społecznej lub reputacja członka grupy społecznej. W środowisku teleinformatycznym pojęcia zaufania lub reputacji są wykorzystywane jako miary rzetelności agentów. Miara zaufania do agenta może być uzależniona od historii interakcji z nim, ale także od rekomendacji innych agentów. Miary zaufania lub reputacji wspomagają podjęcie decyzji dotyczącej wyboru uczestnika interakcji z grupy potencjalnych usługodawców, wśród których mogą znajdować się agenty działające samolubnie lub złośliwie. Wobec tego, zastosowanie systemów TRM pozwala na zredukowanie ryzyka w interakcjach pomiędzy autonomicznymi agentami [1], przyczyniając się do zwiększenia poziomu bezpieczeństwa.

Interdyscyplinarny charakter mechanizmów miękkiego bezpieczeństwa, a także dynamicznie rozwijające się obszary zastosowań systemów TRM, jak również wzrost istotności bezpieczeństwa środowisk teleinformatycznych, wzbudziły wzmożone zainteresowanie autora tą tematyką. Szczególnie interesujące zagadnienie, zdaniem autora niniejszej rozprawy, stanowi kwestia podatności tego typu systemów na specyficzne dla nich ataki, których przeprowadzenie może skutkować zmniejszeniem efektywności działania środowiska lub istotnym ograniczeniem możliwości funkcjonowania niektórych rzetelnych agentów.

Systemy TRM zyskały znaczące zainteresowanie, o czym świadczy bogactwo literatury dotyczącej tego zagadnienia. Dotychczas badania w zakresie systemów TRM koncentrowały się na:

- propozycjach nowych systemów i badaniu ich odporności na wybrane ataki,
- klasyfikacji systemów TRM,
- opisie wybranych typów ataków.

Ze względu na dynamiczny rozwój nowych koncepcji związanych z systemami TRM i mimo wszystko stosunkowo niedługą ich historię, nie została wyczerpująco zbadana odporność tego typu systemów na specyficzne dla nich ataki [1].

Do problemów związanych z systemami TRM, które są słabiej eksplorowane należą:

- opracowanie ogólnego modelu ataków na systemy TRM, umożliwiającego identyfikację nowych form ataków, specyficznych dla danego systemu TRM,
- opracowanie metodyki oceny odporności systemów TRM na ataki, umożliwiającej ich porównywanie w sposób wyczerpujący i usystematyzowany.

Systemy TRM powstały w celu ochrony przed nierzetelnymi agentami, ale mogą być narażone na specyficzne dla nich ataki. Niniejsza praca skupia się na ocenie odporności systemów TRM w kontekście ataków złośliwych, wymierzonych bezpośrednio w same systemy TRM. Wobec tego głównym przedmiotem zainteresowania rozprawy jest to czy systemy zarządzania zaufaniem i reputacją dostarczają wiarygodne informacje, istotnie wspomagające podejmowanie decyzji w warunkach niepewności, nawet przy aktywnym i wyrafinowanym przeciwdziałaniu potencjalnych adwersarzy.

Wskutek braku szeroko akceptowanej metodyki oceny odporności na ataki, poszczególne propozycje systemów TRM są badane w oparciu o odmienne założenia dotyczące sposobu przeprowadzania ataku i z uwzględnieniem różnych miar, co utrudnia porównywanie uzyskiwanych wyników. Opracowanie metodyki oceny odporności na ataki może przyczynić się do bardziej efektywnego badania systemów TRM.

Opracowana w ramach pracy metodyka wykorzystuje, ale nie ogranicza się do, podejścia praktycznego, stosowanego dotychczas w dziedzinie systemów zarządzania zaufaniem i reputacją, w którym analiza wiarygodności koncentruje się na empirycznych badaniach wykrywalności znanych metod ataków. W pracy pokazano także, że przeprowadzenie badań przykładowego systemu TRM w oparciu o stworzoną metodykę jest możliwe i pozwala na zidentyfikowanie konkretnych zagrożeń, z którymi związane jest wykorzystanie danego systemu TRM w określonym środowisku.

1.2. TEZA, CELE I ZAKRES PRACY

Głównym celem rozprawy jest zaproponowanie metodyki oceny odporności systemów TRM na ataki, a teza pracy została sformułowana następująco:

Metodyka oceny wiarygodności systemów zarządzania zaufaniem i reputacją umożliwia dokonanie jakościowej i ilościowej ewaluacji odporności tych systemów na ataki mające za cel zmanipulowanie generowanych wyników i podejmowanych decyzji. Stworzenie metodyki oceny wiarygodności jest możliwe w oparciu o opracowanie modelu środowiska, systemu zarządzania zaufaniem i reputacją oraz generycznego modelu ataku przeciwko tym systemom.

Aby udowodnić przedstawioną tezę i zrealizować powyższy cel główny, praca realizuje cele szczegółowe, do których należą:

- opracowanie modelu środowiska, w którym może być wykorzystywany system TRM;
- opracowanie, w oparciu o istniejące prace, ale w znacznym stopniu je rozszerzając, ogólnego, generycznego modelu systemów TRM;
- opracowanie modelu ataków na systemy TRM oraz kryteriów oceny ich skuteczności w oparciu o generyczny model systemów TRM;
- zaproponowanie metody pozwalającej na znalezienie najbardziej efektywnego ataku na określony system TRM;
- dokonanie opisu przykładowego systemu TRM w ramach opracowanego modelu oraz przeprowadzenie badań jego wiarygodności w oparciu o stworzoną metodykę.

1.3. STRUKTURA PRACY

Rozdział drugi stanowi wprowadzenie do tematyki systemów TRM, poprzez przedstawienie definicji, zastosowań i wybranych cech systemów TRM, istotnych z punktu widzenia osiągnięcia celów rozprawy.

W rozdziale trzecim zawarto przegląd literatury związanej z zagadnieniem ataków i oceną wiarygodności systemów TRM.

Rozdział czwarty zawiera opis modelu środowiska oraz systemu TRM, a także przedstawia modele ataków. Rozdział ten jest szczególnie istotny dla osiągnięcia głównego celu pracy – opracowania metodyki oceny wiarygodności systemów TRM, która została przedstawiona w rozdziale piątym, zawierającym także definicje miar wiarygodności systemów TRM.

Rozdział szósty zawiera badania wybranego systemu TRM w oparciu o opracowaną metodykę, co stanowi przykład jej praktycznego wykorzystania.

Rozdział siódmy zawiera podsumowanie i wnioski z wykonanych prac, a także identyfikację kolejnych istotnych tematów badawczych związanych z systemami TRM.

2. SYSTEMY ZARZĄDZANIA ZAUFANIEM I REPUTACJĄ

Rozdział prezentuje wprowadzenie do tematyki ataków na systemy teleinformatyczne oraz służy zaprezentowaniu podstawowych informacji o systemach TRM. Informacje te są przydatne w kontekście dalszej części pracy, w szczególności w celu dokonania pogłębionego przeglądu literatury, zawartego w następnym rozdziale.

W ramach rozdziału przedstawiono model systemu teleinformatycznego w postaci systemu wieloagentowego (podrozdział 2.1), a także ogólną klasyfikację ataków na systemy teleinformatyczne (podrozdział 2.2) oraz mechanizmów bezpieczeństwa (podrozdział 2.3). Kolejne podrozdziały stanowią wprowadzenie pojęć związanych z jednym z rodzajów mechanizmów bezpieczeństwa systemów teleinformatycznych tj. systemów zarządzania zaufaniem i reputacją. W podrozdziale 2.4 wprowadzono pojęcia zaufania i reputacji, co pozwoliło na zdefiniowanie pojęcia systemu zarządzania zaufaniem i reputacją (w podrozdziale 2.5), które mogą być stosowane w ramach systemu wieloagentowego. Następnie przedstawiono wybrane cechy systemów TRM (w podrozdziale 2.6) oraz przykłady ich zastosowań (w podrozdziale 2.7). Podrozdział 2.8 przedstawia ogólną klasyfikację ataków na systemy TRM, które stanowią specyficzny typ ataków na systemy teleinformatyczne, omawianych wcześniej w podrozdziale 2.2. Ostatni podrozdział – 2.9, przedstawia pojęcie wiarygodności systemu TRM, które jest głównym przedmiotem zainteresowania rozprawy.

2.1. MODEL SYSTEMU TELEINFORMATYCZNEGO

Modelem systemu lub sieci teleinformatycznej może być system wieloagentowy, w którym zbiorowość agentów może nawiązywać interakcje w celu wymiany usług. W zależności od konkretnego systemu teleinformatycznego, zarówno pojęcia agenta, interakcji, jak również usługi mogą być specyficznie definiowane. W przypadku sieci peer-to-peer (P2P), agent może być utożsamiany z użytkownikiem lub klientem sieci P2P, interakcją jest żądanie i wymiana plików, a usługą – oferowanie plików do pobrania. W przypadku sieci ad hoc, agent może być utożsamiany z węzłem sieci, interakcja – z wymianą żądań pomiędzy agentami, a usługa – z przekazaniem pakietu. W przypadku platform e-handlu, agent jest użytkownikiem (klientem), usługą jest dowolny rodzaj produktu oferowanego na sprzedaż, a interakcją – cały proces sprzedaży tego produktu.

W pracy przyjęto następującą definicję systemu wieloagentowego [2]:

Definicja 1: ***System wieloagentowy*** *(ang. „multi-agent system") to system złożony z komunikujących się i współpracujących między sobą agentów, realizujących określone cele.*

W celu uwzględnienia pojęcia interakcji oraz usługi autor zdecydował się na wprowadzenie pojęcia środowiska:

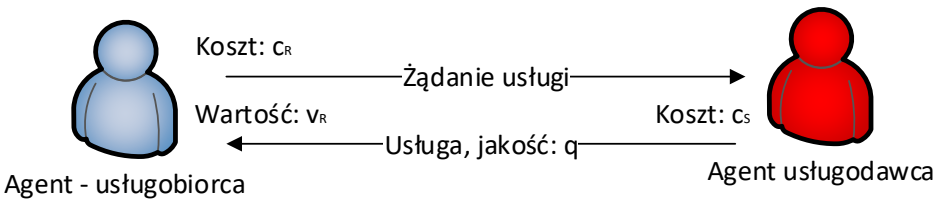
Definicja 2: ***Środowisko*** *to system wieloagentowy, składający się z agentów wchodzących w interakcje polegające na świadczeniu określonych usług.*

W ramach systemu wieloagentowego, złożonego z autonomicznie działających agentów, agenty wchodzą ze sobą w interakcje. Agenty biorące udział w interakcji są określane jako partnerzy interakcji. W ramach interakcji zachodzi świadczenie usługi, przez jednego (lub kilku) z partnerów interakcji. W czasie interakcji, agent (usługobiorca) zgłasza żądanie świadczenia określonej usługi do innego agenta – usługodawcy. W ramach środowiska może być świadczonych wiele usług. Dany agent może żądać określonych typów usług, jak i świadczyć je – jest to zależne od charakterystyki konkretnego agenta. Wystosowanie żądania usługi wiąże się z pewnym kosztem – c_R, który musi zostać poniesiony przez usługobiorcę. Koszt żądania można utożsamiać z zapłatą za wyświadczenie usługi[1]. Podobnie świadczenie usługi przez usługodawcę wiąże się z pewnym kosztem c_S, który jest zależny od jakości świadczonej usługi – q. Wartość usługi v_R, postrzegana przez usługobiorcę jest także zależna od jakości usługi. Warto zauważyć, że w takim modelu, interakcja jest korzystna dla usługobiorcy, jeżeli spełniony jest warunek: $v_R > c_R$, natomiast jest korzystna dla usługodawcy jeżeli jest spełniony warunek[2]: $c_S < c_R$. Wymiar zarówno kosztu żądania, kosztu świadczenia usługi, jak i wartości usługi warunkowanej oceną jej jakości, jest zależny od konkretnego środowiska, np. w systemach e-handlu ma on wymiar pieniądza, w sieciach WSN (ang. Wireless Sensor Networks) zużytego pasma, mocy obliczeniowej lub czasu. Model kosztów interakcji zaprezentowano na rysunku 1. Istotnym problemem z perspektywy usługobiorcy jest to, że jakość usługi nie może być oceniona przed jej wyświadczeniem, co uzasadnia potrzebę

[1] W celu uproszczenia przyjęto założenie, że koszt ponoszony przez żądającego usługę stanowi w całości zapłatę za świadczenie usługi, która trafia do usługodawcy. W niektórych środowiskach to nie musi być spełnione, albo ze względu na to, że istnieją dodatkowe koszty wynikające z funkcjonowania w danym środowisku, albo ze względu na to, że bezpośrednio nie jest możliwe przekazywanie opłat do świadczącego usługę, ale sama możliwość działania w środowisku jest pewnego rodzaju formą zapłaty. Warto także zwrócić uwagę na fakt, że w przypadku niektórych środowisk sam koszt żądania czy świadczenia usługi nie jest tak ważny jak koszt związany z wykorzystaniem nierzetelnej informacji dostarczonej przez agenta (w tym wypadku usługą jest dostarczenie pewnej informacji).

[2] Por. przypis 1

opracowania mechanizmu pozwalającego na ocenę rzetelności agenta – usługodawcy[3]. Użycie mechanizmów oceny zaufania do agenta świadczącego usługę lub reputacji tego agenta pozwala na ocenę jego rzetelności i w ten sposób antycypowanie a priori jakości dostarczanej przez niego usługi, co może przyczynić się do ograniczenia pokusy nadużycia (oszustwa) przez agenta świadczącego usługę i promować działania agentów rzetelnych.



Rysunek 1 Model kosztów interakcji

Na użytek dalszej części pracy warto odnotować dwie uwagi dotyczące interakcji:

Uwaga 2.1: ***Interakcja*** *obejmuje żądanie usługi przez pewnego agenta, oraz jej wyświadczenie na rzecz innego agenta*[4].

Agent działający w środowisku cechuje się następującymi właściwościami [2]:

- **autonomia**: agent podejmuje niezależne od innych agentów decyzje, które są przynajmniej częściowo racjonalne (ich celem jest maksymalizacja własnej użyteczności z funkcjonowania w środowisku);
- **lokalny widok**: żaden agent nie posiada pełnego globalnego obrazu środowiska, lub pełnej wiedzy o środowisku, lub środowisko jest zbyt złożone, aby agent był w stanie wykorzystać w sposób praktyczny taką wiedzę;
- **decentralizacja**: nie ma wyznaczonego agenta, który pełniłby funkcje kontrolne, lub agenty w środowisku są homogeniczne.

Warto podkreślić, że w ramach środowiska mogą działać różnorodne mechanizmy bezpieczeństwa, między innymi w środowisku może działać system zarządzania zaufaniem i reputacją (pojęcie to wprowadzono w podrozdziale 2.5). Przykłady środowisk (systemów wieloagentowych), w których są stosowane systemy zarządzania zaufaniem i reputacją, łącznie z ogólnym określeniem charakterystyki agentów, zostały opisane w podrozdziale 2.7.

[3] Podobnie, w przypadku gdy zapłata za usługę c_R nie trafia do usługodawcy przed jej wyświadczeniem, korzystne dla usługodawcy byłoby posiadanie mechanizmu oceny rzetelności usługobiorcy.
[4] Przyjmuje się, że agent nie może świadczyć usługi samemu sobie, lub jest to pomijane z obszaru rozważań z uwagi na to że jest to trywialny przypadek.

2.2. ATAKI W SYSTEMACH TELEINFORMATYCZNYCH

Ataki przeciwko sieciom i systemom teleinformatycznym (w szczególności dokonywane w sieciach wykorzystujących protokoły IPv4 lub IPv6) mogą być klasyfikowane w oparciu o tak zwaną „piramidę bólu" (ang. „pyramid of pain") [3], co obrazuje rysunek 2. Ideą takiej klasyfikacji, a jednocześnie genezą jej nazwy, jest określenie poziomu trudności związanego z wykryciem danego typu ataku i ewentualnie z przeciwdziałaniem mu. W niniejszym podrozdziale, idea piramidy bólu została jedynie ogólnie zaprezentowana w celu zaznaczenia, że podobne podejście może być zastosowane w kontekście ataków na systemy zarządzania zaufaniem i reputacją.

Rysunek 2 Piramida bólu – wykrywania ataków na systemy teleinformatyczne

Na potrzeby wykrywania ataków oraz wymiany informacji użytecznej (ang. „actionable information") o atakach teleinformatycznych używa się tzw. wskaźników kompromitacji (ang. „Indicator of Compromise" – IoC), które są pewnego rodzaju formą opisu ataku, pozwalającą na jego wykrycie i, na tej podstawie, stwierdzenie czy mogło dojść do kompromitacji danego systemu. IoC to swoista sygnatura ataku. Najłatwiejsze do wykrycia, a co za tym idzie, przeciwdziałania, są ataki, które można opisać poprzez pewną wartość skrótu (np. wartość funkcji skrótu obliczoną z kodu złośliwego oprogramowania). Nieco trudniejsze do pozyskania i przetwarzania (a w szczególności do ograniczenia wyników fałszywie pozytywnych – ang. „false positives") są informacje o atakach, które mogą być opisane za pomocą adresów IP atakującego (np. serwera C&C – ang. „command and control", zarządzającego siecią botnetów). Z jeszcze większą trudnością wiąże się wykrywanie ataków na podstawie nazw domenowych, które są używane w złośliwej komunikacji. Wiedza o zwiększonej trudności

pozyskania użytecznej informacji na cele wykrycia takich ataków jest dość powszechnie wykorzystywana przez twórców złośliwego oprogramowania, którzy używają tzw. algorytmów DGA (ang. „Domain Generation Algorithm"), służących nawiązaniu złośliwej komunikacji bez użycia statycznych nazw domenowych. Kolejnym poziomem piramidy bólu, związanym ze wzrostem trudności pozyskania i przetwarzania informacji użytecznych, są tzw. artefakty hosta (np. obecność pewnych wpisów w rejestrze systemowym Windows, czy obecność określonych plików konfiguracyjnych w systemach Unix i Linux) lub artefakty sieciowe (obecność w komunikacji sieciowej specyficznych pakietów IP, czy komunikacji w oparciu o pewne specyficzne protokoły). Kolejny poziom piramidy bólu dotyczy wykrywania określonych narzędzi służących do przeprowadzania ataku, bazując na charakterystyce ich działania. Wreszcie ostatni poziom dotyczy wszelkiego rodzaju taktyk, technik i procedur (ang. „Tactics, Techniques and Procedures" – TTP) wykorzystywanych przez atakujących do przeprowadzania ataków. W takim przypadku dany atak może zostać scharakteryzowany zwykle jedynie w sposób ogólny poprzez opis zachowania atakujących, a nie narzędzi wykorzystywanych do ataku, czy innych wskaźników, które zostały uwzględnione na niższych poziomach piramidy. Warto zwrócić uwagę, że na im wyższym poziomie piramidy odbywa się reakcja na atak, tym jest to efektywniejsze (zdecydowanie trudniej zmienić atakującemu sposób zachowania, niż narzędzie wykorzystywane do tego celu, a tym bardziej adres IP atakującego). Także opis ataku w sposób umożliwiający automatyczne wykrywanie ataku jest tym trudniejszy im wyższy poziom piramidy. W szczególności, na wyższych poziomach piramidy, opis zachowania atakującego jest zagadnieniem wysoce nietrywialnym, a zasoby potrzebne do wykonywania monitorowania pojawiających się wskaźników kompromitacji są znacząco większe niż na niższych poziomach piramidy.

Powyższy opis idei piramidy bólu dla ataków na sieci i systemy teleinformatyczne jest daleki od wyczerpującego, szerszy opis może być znaleziony w literaturze [3]. Idea ta jest powszechnie wykorzystywana w działalności operacyjnej zespołów reagowania na incydenty komputerowe (CSIRT – ang. „Computer Security Incident Response Team").

W myśl tak określonej, ogólnej (przeznaczonej dla szerokiego spektrum ataków na systemy teleinformatyczne, ze szczególnym uwzględnieniem sieci IP) piramidy bólu, wszystkie ataki przeciwko systemom TRM są dokonywane na poziomie TTP. Interesującym zagadnieniem jest stworzenie dedykowanej piramidy bólu w odniesieniu do systemów TRM, co zostało wykonane w dalszej części pracy, w podrozdziale 2.8.

2.3. MECHANIZMY BEZPIECZEŃSTWA W SYSTEMACH TELEINFORMATYCZNYCH

W systemach teleinformatycznych mogą być stosowane dwa typy mechanizmów bezpieczeństwa:

- mechanizmy twardego bezpieczeństwa,
- mechanizmy miękkiego bezpieczeństwa.

Mechanizmy twardego bezpieczeństwa jako główny cel mają za zadanie nie dopuścić do uzyskania dostępu do systemu lub informacji w nim przetwarzanych przez agenty, które są nierzetelne. Mechanizmy twardego bezpieczeństwa najczęściej wykorzystują kryptografię, a można je zdefiniować następująco [4], [5]:

Definicja 3: ***Mechanizmy twardego bezpieczeństwa*** *(ang. „hard security") – są związane z indywidualnymi metodami ochrony, których skuteczność opiera się na ściśle określonych regułach działania i ugruntowanych podstawach matematycznych.*

Mechanizmy miękkiego bezpieczeństwa nie mają na celu niedopuszczenia nierzetelnych agentów do funkcjonowania w systemie, ponieważ z wielu względów może to być niemożliwe, na przykład z uwagi na fakt, że system musi być otwarty. Mechanizmy te skupiają się na identyfikacji takich agentów i zapobieganiu ich niekorzystnemu wpływowi na efektywność działania systemu. Mechanizmy miękkiego bezpieczeństwa mogą być zdefiniowane następująco [4], [5]:

Definicja 4: ***Mechanizmy miękkiego bezpieczeństwa*** *(ang. „soft security") – są związane z kolektywnymi metodami ochrony, które zakładają, że w systemie może znaleźć się intruz lub nieoczekiwanie działający komponent. Zadaniem mechanizmów miękkiego bezpieczeństwa jest wykrycie tego rodzaju zdarzeń lub adwersarzy i uniemożliwienie im spowodowania szkód dla całego systemu.*

Przykładem mechanizmów miękkiego bezpieczeństwa są systemy zarządzania zaufaniem i reputacją, a także wszelkie mechanizmy, które wykorzystują świadomość kontekstu. W związku z tym, główny przedmiot pracy stanowi podzbiór mechanizmów miękkiego bezpieczeństwa w postaci systemów zarządzania zaufaniem i reputacją, a w szczególności metody oceny odporności tego typu systemów na ataki.

2.4. POJĘCIA ZAUFANIA I REPUTACJI

Pojęcia zaufania i reputacji zarówno w literaturze dotyczącej systemów zarządzania zaufaniem i reputacją, jak i w ogólnie stosowanym języku, są definiowane na wiele sposobów. Słownik języka polskiego w następujący sposób określa reputację, zaufanie i ufność:

- **reputacja** – *„opinia, jaką ktoś lub coś ma wśród ludzi”*[6], lub *„opinia, renoma, dobre imię, rozgłos, sława”*[7];
- **zaufanie** – *„przeświadczenie, że komuś można ufać, ufność”*[6], lub:
  „1. «przekonanie, że jakiejś osobie lub instytucji można ufać»
  «2. przekonanie, że czyjeś słowa, informacje itp. są prawdziwe»
  «3. przekonanie, że ktoś posiada jakieś umiejętności i potrafi je odpowiednio wykorzystać»”[7];
- **ufność** – *„przeświadczenie, że komuś, czemuś można ufać, wierzyć, że można polegać na danej osobie, zaufanie”* [6].

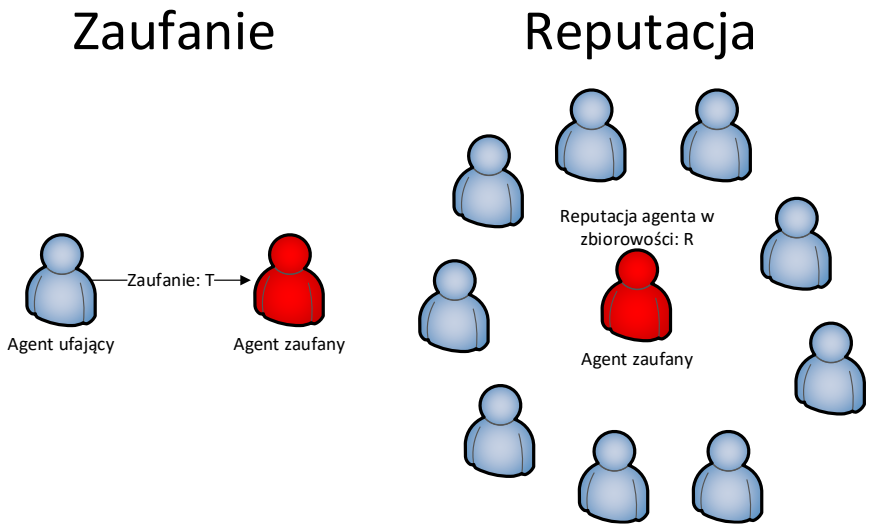
Powyższe definicje niebezpośrednio odnoszą się do istoty działania systemów zarządzania zaufaniem i reputacją.

W literaturze dotyczącej systemów TRM są proponowane różne definicje zaufania i reputacji, szeroki przegląd literatury pod kątem definicji zaufania został także wykonany, m.in. przez autora publikacji [8]. Na potrzeby tej rozprawy, po analizie literatury, autor zdecydował się na przyjęcie następujących definicji zaufania i reputacji:

Definicja 5: ***Zaufanie*** *charakteryzuje relację pomiędzy parą agentów, dotyczącą określonego kontekstu, będącą oceną agenta ufającego co do rzetelności zachowania agenta zaufanego. Wobec tego* ***relacja zaufania*** *jest uporządkowaną 5-elementową krotką (5-ką)* (A_1, A_2, v, c, m)*, gdzie* A_1 *jest agentem ufającym,* A_2 *agentem zaufanym,* v *jest pewną wartością zaufania z określonego zbioru możliwych wartości, charakteryzującą moc relacji zaufania,* c *charakteryzuje kontekst relacji zaufania, a* m *jest czasem, w którym dana relacja zachodzi.*

Definicja 6: ***Reputacja*** *charakteryzuje opinię o danym agencie, dotyczącą określonego kontekstu, będącą miarą oceny grupy agentów co do rzetelności zachowania tego agenta. Wobec tego* ***opinia o reputacji*** *jest uporządkowaną 4-elementową krotką (4-ką)* (A_2, v, c, m)*, gdzie* A_2 *jest agentem zaufanym,* v *jest pewną wartością reputacji z określonego zbioru możliwych wartości, charakteryzującą poziom reputacji agenta,* c *charakteryzuje kontekst reputacji, a* m *jest czasem, w którym dana relacja zachodzi.*

Warto podkreślić, że zaufanie charakteryzuje relację pomiędzy dwoma agentami, natomiast reputacja charakteryzuje konkretnego agenta w zbiorowości innych agentów. Schematycznie relację zaufania i reputacji przedstawiono na rysunku 3.



Rysunek 3 Zaufanie a reputacja

Można wyróżnić następujące własności relacji zaufania:

- **częściowa przechodniość** – własność może być wyrażona słownie w następujący sposób: *jeżeli agent A ufa agentowi B i agent B ufa agentowi C, to agent A może przynajmniej częściowo ufać agentowi C*, tzn.:

  $(A_A, A_B, v_{AB}, c, m) \wedge (A_B, A_C, v_{BC}, c, m) => (A_A, A_C, v_{AC}, c, m)$, przy czym v_{AC} jest większe od wartości minimalnej zaufania (czyli została nawiązana pewna relacja zaufania pomiędzy A i C)

- **częściowa addytywność** – własność może być wyrażona słownie w następujący sposób: *jeżeli agent A ufa agentom B i C oraz agent B ufa agentowi D, to agent A będzie ufał bardziej agentowi D jeżeli dodatkowo agent C ufa agentowi D*, tzn. z warunków:

$$(A_A, A_B, v_{AB}, c, m) \wedge (A_A, A_C, v_{AC}, c, m) \wedge (A_B, A_D, v_{BD}, c, m) => (A_A, A_D, v_{AD}, c, m);$$
$$(A_A, A_B, v_{AB}, c, m) \wedge (A_A, A_C, v_{AC}, c, m) \wedge (A_B, A_D, v_{BD}, c, m) \wedge (A_C, A_D, v_{CD}, c, m) => (A_A, A_D, v'_{AD}, c, m)$$

  wynika że: $v'_{AD} \geq v_{AD}$

- **niesymetryczność (kierunkowość)** – własność może być wyrażona słownie w następujący sposób*: z faktu, że agent A ufa agentowi B nie wynika, że agent B ufa agentowi A*, tzn.:

  $\sim \big((A_A, A_B, v_{AB}, c, m) \wedge (A_B, A_A, v_{BA}, c, m) => v_{AB} = v_{BC}\big)$.

Idea stojąca za stosowaniem pojęć zaufania i reputacji w celu zwiększenia bezpieczeństwa systemów teleinformatycznych, może być wyrażona poprzez własność, znaną z interakcji społecznych:

Własność 2.1*: Im wyższa wartość zaufania do danego agenta lub im wyższa wartość reputacji danego agenta w określonym kontekście, tym mniejsze jest ryzyko związane z zachowaniem tego agenta w tym kontekście.*

Ryzyko może być definiowane np. w sposób dobrze znany z podejścia prezentowanego w przypadku oceny bezpieczeństwa systemów teleinformatycznych, jako wypadkowa prawdopodobieństwa zaistnienia określonego (niepożądanego) zdarzenia i jego skutków (rozumianych np. jako poniesione koszty interakcji).

W nieco odmiennym ujęciu, przypuszczalnie najczęściej stosowanym, pomijany jest aspekt skutków nierzetelnego działania danego agenta jako uczestnika określonej interakcji. W związku z czym zaufanie lub reputacja pełni funkcję oszacowania wartości prawdopodobieństwa nierzetelnego działania agenta.

Własność 2.2*: Im wyższa wartość zaufania do danego agenta lub im wyższa wartość reputacji danego agenta w określonym kontekście, tym niższe prawdopodobieństwo nierzetelnego zachowania tego agenta w danym kontekście (np. świadczenia usługi niskiej jakości).*

2.5. SYSTEMY ZARZĄDZANIA ZAUFANIEM I REPUTACJĄ

W odniesieniu do systemów zarządzania zaufaniem i reputacją, w literaturze anglojęzycznej funkcjonują określenia: „Trust and Reputation Systems”, „Trust and Reputation Management Systems”; szczególnie często jest również używane określenie „model” (np. ang. „Trust and Reputation Model”, „Trust Model”, „Reputation Model”, „Trust Management Model”, „Trust and Reputation Management Model”). Brak jednolitego aparatu pojęciowego prowadzi do daleko idących niejednoznaczności.

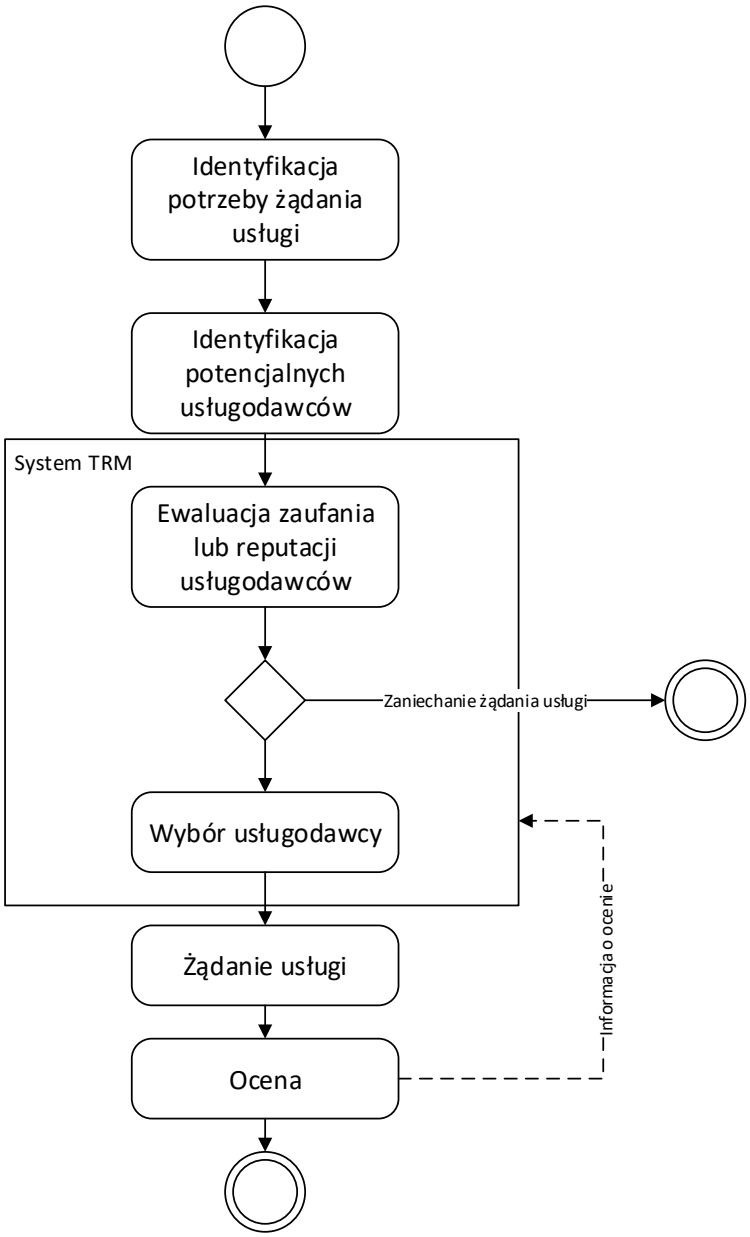
W literaturze stosowanych jest wiele definicji systemu zarządzania zaufaniem i reputacją (np. [1], [9]–[12]). Według publikacji [8] zarządzanie zaufaniem jest obszarem techniki informacyjnej, który ma na celu polepszenie działania otwartych, rozproszonych systemów poprzez przewidywanie lub wpływanie na zachowanie użytkowników, a systemy zarządzania zaufaniem mogą zostać zastosowane do kontroli zachowania agentów.

Na potrzeby rozprawy, warto przytoczyć najczęściej wskazywane w literaturze właściwości systemów TRM, które wskazują, że system zarządzania zaufaniem i reputacją to system złożony z komunikujących i współpracujących między sobą agentów, realizujących określone cele, definiujący zbiór ustalonych reguł podejmowania decyzji przez agenty o nawiązaniu interakcji z innymi agentami lub wyboru konkretnego agenta jako usługodawcy, określających sposób obliczania miar zaufania lub reputacji, charakteryzujących innych agentów oraz wymiany informacji pomiędzy agentami.

Na podstawie analizy literatury, autor pracy zdecydował się na wprowadzenie następującej opisowej definicji systemu TRM:

Definicja 7: ***System zarządzania zaufaniem i reputacją (system TRM)*** *to system, działający w ramach określonego środowiska, który na bazie informacji o ocenie rzetelności wymienianej pomiędzy agentami lub na bazie informacji o ocenie rzetelności dostarczanych do i przez zaufaną trzecią stronę, wspomaga decyzję agenta co do wyboru przyszłego partnera interakcji lub jakości świadczonej usługi, w celu maksymalizacji własnej użyteczności agenta lub środowiska.*

Rysunek 4 prezentuje schematycznie całościowy proces świadczenia usługi, ze wskazaniem wpływu systemu TRM na ten proces.



Rysunek 4 Diagram aktywności w procesie świadczenia usługi ze wskazaniem zakresu działania systemu TRM

Systemy zarządzania zaufaniem i reputacją mogą jednocześnie wykorzystywać zarówno miary zaufania jak i reputacji, ale najczęściej używają tylko miary zaufania albo tylko miary reputacji, przez co czasami stosowany jest podział na systemy zarządzania zaufaniem (systemy zaufania – ang. „trust systems" lub „trust models") i systemy reputacyjne (ang. "reputation systems" lub "reputation management systems/models"). Podział ten bywa jednak nieścisły i nie zawsze jest konsekwentnie stosowany. Z tego względu, a także z uwagi na przeważające podobieństwa systemów wykorzystujących pojęcia zaufania i reputacji w stosunku do różnic pomiędzy nimi, autor zdecydował się na ich łączne potraktowanie, szczególnie, że takie podejście zdaje się przeważać w literaturze.

Celem działania systemów TRM jest ochrona przed atakami na działanie danego środowiska i maksymalizacja użyteczności z wykorzystania środowiska oraz swoiste zachęcenie do poprawy jakości świadczonych usług przez poszczególnych agentów.

Problem uzasadniający użycie systemów TRM może został opisany przez autora rozprawy we wcześniejszej publikacji [13] w następujący sposób: „*W danym środowisku istnieje wiele agentów, funkcjonujących w sposób autonomiczny (niezależny). Agenty są w stanie świadczyć usługę (lub usługi) o określonej jakości, której dostarczenie wiąże się z kosztem*[5] *uzależnionym od typu oraz jakości świadczonej usługi. Część z agentów w danym środowisku jest rzetelna, tzn. że świadczą rzetelnie usługi, ale występują też agenty samolubne – mające na celu maksymalizację własnych zysków z uczestnictwa w środowisku, przy minimalizacji kosztu i agenty złośliwe – dążące do zaburzenia efektywności funkcjonowania środowiska. Wobec tego, w momencie potrzeby skorzystania z usługi, agent natrafia na problem decyzyjny dotyczący wyboru usługodawcy*[6]*. Ze względu na to, że samo żądanie usługi (i późniejsze przeprowadzenie interakcji) wiąże się z pewnym kosztem, agent żądający powinien postępować w sposób, który pozwoli na wybranie jako uczestnika interakcji agenta rzetelnego oraz uniknąć interakcji z agentami samolubnymi i złośliwymi. Systemy zarządzania zaufaniem i reputacją mają wspomagać ten proces decyzyjny.”*

Funkcjonowanie systemów TRM jest uzasadnione w przypadku gdy:

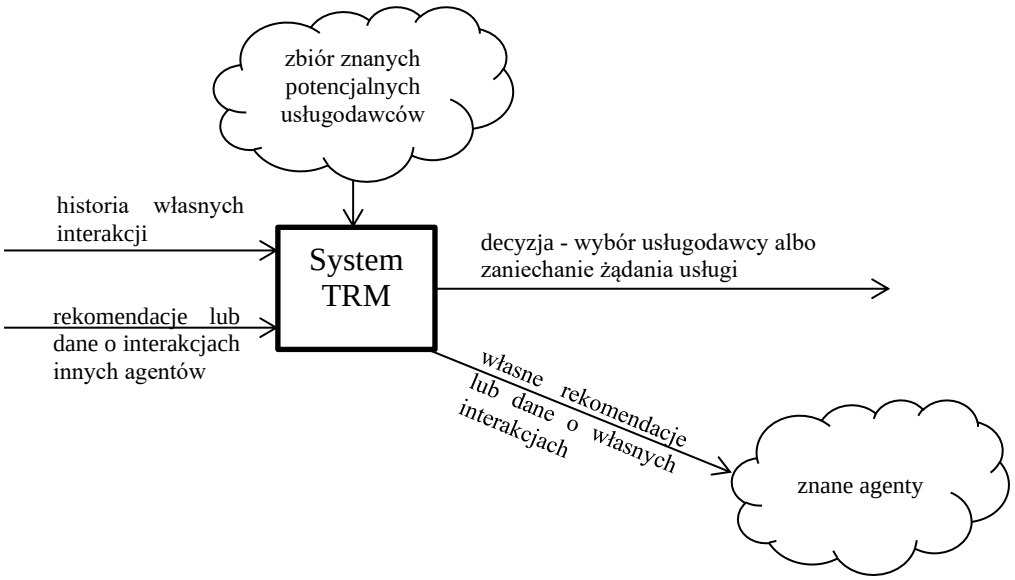
- usługę może świadczyć kilku usługodawców – istnieje wtedy problem decyzyjny polegający na wyborze usługodawcy;
- istnieje tylko jeden usługodawca w danym środowisku, ale jest możliwość podjęcia decyzji czy skorzystać z usługi, czy też zrezygnować z jej żądania, ze względu np. na znaczne ryzyko niesatysfakcjonującej jakości tej usługi;
- poszczególne agenty dostarczają niespójnych rekomendacji lub istnieje potrzeba oceny jakości dostarczonych rekomendacji.

Idea funkcjonowania systemów TRM została przedstawiona na rysunku 5, przytaczając artykuł [13] autora rozprawy, można stwierdzić, że: *„w oparciu o historię własnych interakcji oraz otrzymywane rekomendacje pochodzące od innych agentów, system TRM podejmuje*

[5] Definicja pojęcia kosztu jest uzależniona od charakterystyki środowiska w jakim funkcjonuje system zarządzania zaufaniem i reputacją, np. w sieciach WSN koszt może być utożsamiany z wykorzystaniem mocy obliczeniowej, wykorzystaniem pasma lub zużyciem energii na cele przekazywania (routingu) pakietu; w platformach e-commerce jest to koszt nabycia danego produktu.

[6] Przy założeniu, że wystarczy aby usługę dostarczył tylko jeden agent oraz, że istnieje więcej niż jeden potencjalny usługodawca, co w ogólnym przypadku nie zawsze jest spełnione. Wtedy problem decyzyjny może obejmować także rozstrzygnięcie, czy ze względu na potencjalne ryzyko związane z interakcją, agent nie powinien zrezygnować z tej interakcji.

decyzję, który z agentów ze zbioru potencjalnych usługodawców[7] powinien zostać wybrany jako usługodawca. (...) Na bazie systemu TRM analizowany agent jest w stanie wydawać także rekomendacje lub przesyłać dane o własnych interakcjach do innych agentów, które mogą je wykorzystać do podejmowania własnych decyzji co do wyboru agenta jako usługodawcy. Możliwe jest także wykorzystanie systemów TRM do podjęcia decyzji przez usługodawcę, co do tego o jakiej jakości usługę świadczyć dla agenta żądającego usługi[8]."



Rysunek 5 Wybór usługodawcy (partnera interakcji) przez agenta żądającego usługę w oparciu o system TRM, opracowanie własne na podstawie [13]

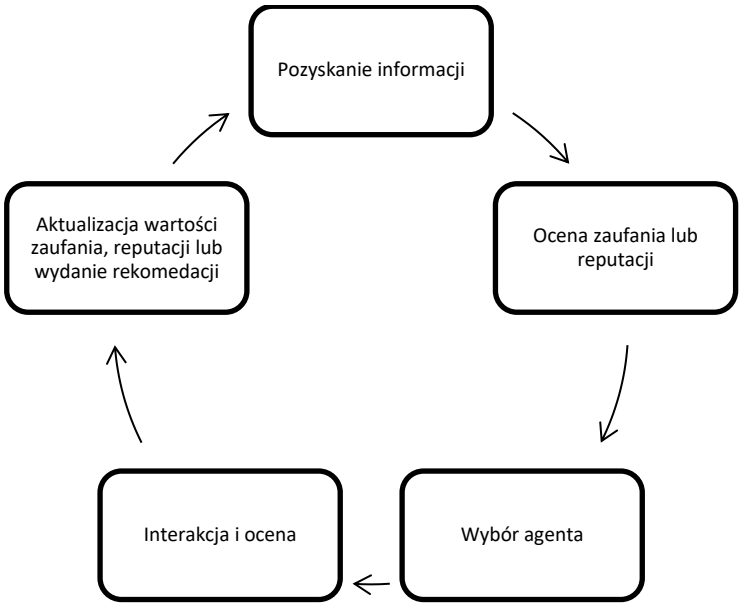
Systemy TRM mogą wykorzystywać rekomendacje otrzymane od innych agentów lub informacje o interakcjach innych agentów. Rekomendacja może być rozumiana jako wskazówka dla agenta otrzymującego rekomendację co do tego, czy agent, którego dotyczy rekomendacja, jest przez agenta wydającego rekomendację oceniany jako rzetelny. Dane o interakcji obejmują najczęściej oznaczenie uczestników interakcji i jej wynik. W niektórych systemach TRM wykorzystywane są tylko wysokopoziomowe binarne rekomendacje lub w ogóle nie są wykorzystywane informacje pochodzące od innych agentów.

[7] Problem wyznaczenia tych agentów, które mogą być potencjalnymi usługodawcami nie leży w zakresie systemu TRM, ale powinien być rozstrzygnięty w ramach innych mechanizmów środowiska, w którym dany system TRM funkcjonuje (najczęściej agent żądający usługę musi utrzymywać pewnego rodzaju relację z agentami, które zostaną określone jako potencjalni usługodawcy, lub agenty muszą w pewien sposób zgłosić gotowość do świadczenia określonej usługi do poszczególnych agentów lub pewnego rodzaju katalogu usług w danym środowisku).

[8] Możliwość stosowania przez usługodawców dyskryminacji (zróżnicowania) jakości usług w zależności od usługobiorcy jest uzależniona od charakterystyki środowiska i świadczonych usług, niemniej czasami rozważana w kontekście systemów TRM.

Funkcjonowanie systemu TRM w ramach procesu świadczenia usług, można podzielić na następujące cyklicznie etapy [2], zaprezentowane schematycznie na rysunku 6:

1. Pozyskanie informacji, polegające na obserwowaniu interakcji innych agentów, pozyskiwaniu rekomendacji od innych agentów, lub przechowywaniu wyników własnych poprzednich interakcji z innymi agentami;
2. Ocena zaufania do poszczególnych agentów na podstawie zebranych informacji, lub ocena reputacji poszczególnych agentów;
3. Wybór agenta świadczącego usługę;
4. Interakcja i ocena jej wyniku;
5. Aktualizacja wartości zaufania lub reputacji, czyli zmniejszenie lub zwiększenie zaufania do agenta lub reputacji agenta świadczącego usługę w zależności od oceny jakości interakcji, etap ten może także obejmować propagację informacji o wyniku interakcji, o zmianie wartości zaufania/reputacji lub o pozytywnej lub negatywnej rekomendacji.



Rysunek 6 Etapy działania systemu TRM

Etapy te mogą być znacząco oddzielone od siebie w czasie, a co więcej mogą być przedzielone wystąpieniem kolejnych cykli działania systemu TRM, mogą także nie wystąpić w ogóle. Na przykład, w przypadku systemów e-handlu, agent (użytkownik) po pozyskaniu informacji, ocenie zaufania i wyborze użytkownika, od którego dokona zakupu, przystępuje do sfinalizowania transakcji – zamawia i opłaca dany towar. Jednakże możliwość oceny interakcji wystąpi dopiero po znaczącym czasie w stosunku do czasu trwania poprzednich etapów – w momencie gdy towar dotrze do odbiorcy (lub po jeszcze dłuższym czasie, gdy okaże się że

sprzedający jest oszustem i nie wysłał towaru). W tym czasie zarówno kupujący jak i sprzedający mogą być stronami innych interakcji (transakcji), jak również w tym środowisku (platformy e-handlu) mogą zachodzić interakcje pomiędzy innymi agentami.

Szczególnie istotną cechą systemów TRM jest to, że stan systemu w ramach danego cyklu zależy od stanów systemu w ramach poprzednich cyklów.[9] Cecha ta jest oczywistą konsekwencją faktu wykorzystania informacji historycznych (bezpośrednio - np. wyników przeszłych interakcji, lub pośrednio – np. rekomendacji, które są wynikiem doświadczenia agentów zdobytego na bazie wyników historycznych interakcji).

2.6. WYBRANE CECHY SYSTEMÓW TRM

Systemy TRM można klasyfikować biorąc pod uwagę wiele cech. Celem podrozdziału nie jest prezentacja dogłębnej klasyfikacji ani przegląd klasyfikacji dokonanych w literaturze, a jedynie zestawienie tych cech systemów TRM, które są szczególnie istotne dla analizy ich wiarygodności, rozumianej jako odporność na ataki. Jednocześnie ogólny przegląd literatury dotyczący zagadnienia klasyfikacji i tworzenia taksonomii systemów TRM został dokonany w rozdziale 3.

W systemie TRM do oceny rzetelności agentów mogą być wykorzystywane parametry dotyczące:

- reputacji – jako scentralizowanej oceny każdego agenta,
- zaufania – jako oceny jednego agenta dokonanej przez innego agenta pomiędzy każdą parą agentów, które nawiązały relację.

Systemy TRM mogą wykorzystywać następujące źródła informacji służące do oceny zaufania lub reputacji (przy czym wiele systemów TRM wykorzystuje więcej niż jedno źródło informacji):

- własne interakcje,
- bezpośrednie obserwacje,
- scentralizowane dane o interakcjach innych agentów,
- dane o interakcjach innych agentów dostarczane przez agenty bezpośrednio zaangażowane w interakcję,
- scentralizowane dane o ocenie reputacji lub zaufania konkretnego agenta,

[9] W przypadku niektórych systemów TRM stany systemu z zamierzchłej przeszłości mogą mieć mniejszą istotność lub w ogóle mogą nie być brane pod uwagę.

- dane o ocenie reputacji lub zaufania konkretnego agenta dostarczane przez innego agenta.

Systemy TRM mogą dokonywać wyznaczenia wartości zaufania lub reputacji w sposób:

- scentralizowany (w systemie istnieje centralna jednostka, która prezentuje wyniki oceny dla poszczególnych agentów; najczęściej ten sposób występuje w systemach wykorzystujących jedynie reputację),
- rozproszony (poszczególne agenty same dokonują kalkulacji zaufania lub reputacji w oparciu o dostarczone informacje; najczęściej ten sposób występuje w systemach wykorzystujących zaufanie).

Obliczone miary zaufania lub reputacji mogą być wykorzystywane do podjęcia decyzji o wyborze partnera interakcji w następujący sposób:

- wybór agenta o najwyższej wartości zaufania lub reputacji – interakcja jest nawiązywana z agentem, który charakteryzuje się najwyższą wartością zaufania lub reputacji;
- wybór agenta w sposób probabilistyczny – nawiązanie interakcji z danym agentem jest uzależnione od wartości zaufania lub reputacji w ten sposób, że im wyższa wartość zaufania lub reputacji tym wyższe prawdopodobieństwo wyboru danego agenta do interakcji;
- wybór agenta w sposób probabilistyczny z określonym progiem – nawiązanie interakcji z danym agentem jest uzależnione od wartości zaufania lub reputacji w ten sposób, że poniżej pewnego progu zaufania lub reputacji nawiązanie interakcji z danym agentem jest niemożliwe, a powyżej tego progu, prawdopodobieństwo nawiązania relacji jest tym większe im większa wartość zaufania lub reputacji tego agenta;
- wybór agenta określony inną funkcją – jest to uogólnienie powyższych przypadków, w którym wybór agenta następuje w sposób określony pewną funkcją, której parametrami są wartości zaufania lub reputacji[10];
- wartości zaufania lub reputacji mają wartość jedynie informacyjną – sposób podejmowania decyzji o nawiązaniu interakcji nie jest częścią specyfikacji systemu lub decyzja jest każdorazowo podejmowana samodzielnie przez agenta danego środowiska w sposób niestandardowy.

[10] W praktyce, najczęściej systemy TRM wykorzystują jeden z trzech powyższych sposobów wyboru agenta, sposób czwarty jest uogólnieniem trzech pierwszych sposobów.

2.7. ZASTOSOWANIA SYSTEMÓW TRM

Systemy zarządzania zaufaniem i reputacją są najczęściej wykorzystywane w różnych typach sieci, a także w platformach e-handlu (portalach aukcyjnych, sklepach internetowych), czy aplikacjach mobilnych. Niniejszy podrozdział przedstawia praktyczne przykłady zastosowań systemów zarządzania zaufaniem i reputacją.

2.7.1. Platformy e-handlu

W przypadku serwisów internetowych oferujących sprzedaż produktów lub usług, systemy zarządzania zaufaniem i reputacją znajdują zastosowanie do ochrony kupujących, rekomendowania produktów lub usług, jak również, rzadziej – do ochrony sprzedających. Propozycje systemów TRM dla platform e-handlu zawarto m.in. w pracach: [14]–[18]. Ogólną cechą większości ze stosowanych systemów TRM jest ich nieskomplikowanie (co można niewątpliwie wskazywać jako ich zaletę), ale też brak uwzględnienia, podczas ich tworzenia, możliwych ataków przeciwko takim systemom. W wielu przypadkach głównym sposobem obrony przed atakami polegającymi na publikowaniu nierzetelnych rekomendacji, jest weryfikacja dokonywana przez system, czy oceniana transakcja (interakcja) miała w rzeczywistości miejsce. Dość częste jest też zastępowanie pierwszych ocen przez pewną średnią ocen lub ich utajnianie dopóki nie pojawi się odpowiednia pula ocen, po to aby pierwsze lub skrajne oceny nie miały zbyt dużego wpływu na decyzje kupujących.

2.7.1.1. Sklepy internetowe

W przypadku sklepów internetowych, pewne, w większości bardzo proste, systemy zarządzania zaufaniem i reputacją są wykorzystywane do oceny oferowanych produktów. W wielu przypadkach wydawane rekomendacje mają postać słownej opinii o danym produkcie, ale dość często jest także wykorzystywana niezłożona ocena ilościowa danego produktu (np. ocena w skali od 1 do 5). Oceny te są następnie agregowane (bardzo często w postaci obliczenia średniej arytmetycznej ocen poszczególnych klientów), stając się pewnego rodzaju „reputacją” – oceną jakości danego produktu.

2.7.1.2. Serwisy agregujące informacje o produktach i sklepach internetowych

W Internecie funkcjonują serwisy agregujące i prezentujące informacje o produktach i przechowujące opinie (rekomendacje) dostarczone przez ich użytkowników. Przykładami takich serwisów są opineo.pl, czy ceneo.pl. W serwisach tych możliwe jest dokonanie oceny rzetelności konkretnego sklepu internetowego (najczęściej w skali od 1 do 5) oraz wystawienie

opisowej oceny wynikającej z własnego doświadczenia. W przypadku tych serwisów, łatwo można dostrzec istnienie ataków, polegających na dostarczaniu nierzetelnych rekomendacji (np. wydawanych bezpośrednio przez producentów na temat własnych produktów, lub produktów konkurencji, czy wydawanych przez właścicieli sklepu internetowego o własnym sklepie) i łatwe zmanipulowanie w ten sposób potencjalnych klientów.

2.7.1.3. Platformy aukcyjne

W przypadku serwisów aukcyjnych, takich jak allegro.pl, czy eBay.com, rzetelność zarówno sprzedających, jak i kupujących jest oceniana przez drugą stronę transakcji (interakcji). Rekomendacja obejmuje najczęściej binarne określenie wyniku interakcji (pozytywny lub negatywny) oraz dodatkową ocenę poszczególnych elementów interakcji w pięciostopniowej skali. W tych serwisach aukcyjnych możliwe jest dokonanie wielu ze zidentyfikowanych ataków na systemy zarządzania zaufaniem i reputacją (np. posiadanie wielu kont przez jednego sprzedającego, wydawanie fałszywych rekomendacji itd.).

2.7.1.4. Serwisy rezerwacji noclegu

W przypadku serwisów umożliwiających rezerwację noclegu, takich jak booking.com, czy airbnb.com, także są wykorzystywane proste systemy zarządzania zaufaniem i reputacją, bazujące na rekomendacjach wydawanych przez klientów, którzy skorzystali z usługi rezerwacji danego miejsca noclegowego. Wystawione i zagregowane oceny (zarówno ilościowe w określonej skali, jak i opisowe) mają umożliwić dokonywanie lepszych decyzji kolejnym klientom i promować te miejsca noclegowe, które cechują się wysoką jakością. Także i w tym przypadku wystawiane opinie mogą być zmanipulowane, na przykład poprzez uniemożliwienie wystawienia negatywnej opinii z uwagi na deklarację hotelu, że do wizyty i transakcji nie doszło.

2.7.1.5. Serwisy w ramach Dark Web

Niektóre z serwisów dostępnych w tzw. Dark Web (ukrytej części Internetu) także wykorzystują systemy zarządzania zaufaniem i reputacją, na przykład w celu uniemożliwienia dostępu do niektórych treści lub ofert dla użytkowników, którzy nie mają wystarczającego poziomu reputacji. Przykładem jest serwis Cebulka[11], który był dostępny w ramach sieci TOR. W wielu przypadkach jest to sposób na utrudnienie infiltracji dokonywanej przez organa ścigania, jako że znacząca część tych serwisów oferuje produkty lub usługi nielegalne.

[11] http://qd73mvvc7v7zewwl.onion/

2.7.2. Aplikacje mobilne

Istnieją także aplikacje mobilne, wykorzystujące w swoim działaniu systemy zarządzania zaufaniem i reputacją. Przykłady dwóch taki aplikacji zaprezentowano w bieżącym punkcie.

2.7.2.1. Aplikacja yanosik

Aplikacja yanosik, oprócz funkcjonalności nawigacji samochodowej, umożliwia także ostrzeganie innych uczestników ruchu przed zagrożeniami na drogach, czy kontrolach prędkości. Każdy z użytkowników aplikacji ma możliwość zaraportowania danego rodzaju zdarzenia drogowego, a informacja o nim trafi do wszystkich innych użytkowników aplikacji, którzy mogą być zainteresowani taką informacją, tzn. do tych, którzy znajdują się w pobliżu miejsca zdarzenia. W celu wykrywania nierzetelnych zgłoszeń, w ramach aplikacji funkcjonuje pewien system zarządzania zaufaniem i reputacją, oceniający rzetelność użytkowników. Według twórców aplikacji [19]: *„Ocena wiarygodności użytkownika (ranga) obliczana jest na podstawie jego aktywności w systemie i rzetelności jego zgłoszeń. Ocena w skali od 0 do 5 gwiazdek nie jest przyznawana na stałe - na wiarygodność trzeba sobie zapracować. Algorytm obliczania wiarygodności bierze pod uwagę różne czynniki, jak np. to czy Twoje zgłoszenia są potwierdzane, czy odwoływane. Ranga wyświetlana jest w aplikacji, na stronie w profilu użytkownika oraz przy postach na forum."* Pojęcie wiarygodności używane przez yanosik.pl odpowiada pojęciu reputacji w ramach tej pracy. Niestety nie jest dostępny dokładniejszy i bardziej formalny opis systemu zarządzania zaufaniem stosowanego w ramach tej aplikacji.

2.7.2.2. Aplikacja odebrać telefon?

Aplikacja „Odebrać telefon?" to narzędzie do walki ze spamem telefonicznym, czyli niechcianymi połączeniami, mającymi za cel marketing bezpośredni pewnych produktów, lub wręcz dokonywanie oszustw lub wyłudzeń. Każdy z użytkowników aplikacji może zgłosić dany numer telefonu jako niebezpieczny lub niechciany. Dzięki temu gdy nastąpi w przyszłości połączenie z takiego potencjalnie niebezpiecznego numeru do któregokolwiek z użytkowników aplikacji, informacja o możliwym spamie zostanie wyświetlona przez aplikację lub połączenie to zostanie automatycznie odrzucone. W ramach aplikacji działa pewien system zarządzania zaufaniem i reputacją mający za zadanie odrzucanie niewłaściwych rekomendacji (opinii o połączeniach), ale opis jego funkcjonowania nie jest udostępniony przez twórców aplikacji.

2.7.3. Repozytoria oprogramowania

Repozytoria oprogramowania, takie jak sklep Google Play z aplikacjami przeznaczonymi na urządzenia z systemem operacyjnym Android, jak również App Store z aplikacjami przeznaczonymi z urządzeniami z systemem operacyjnym iOS, czy Microsoft Store zawierający aplikacje dla systemów operacyjnych Windows, także wykorzystują pewnego rodzaju ocenę udostępnianego oprogramowania, którą można utożsamiać z bardzo prostym systemem zarządzania zaufaniem i reputacją. Oceny te uwzględniają wystawiane przez użytkowników opinie na temat danej aplikacji, jak również jej popularność (liczbę użytkowników).

2.7.4. Internet Rzeczy

Po raz pierwszy pojęcie Internetu Rzeczy (ang. „Internet of Things" – IoT) zostało użyte w 1999 roku przez Kevina Ashtona [20]. Idea Internetu Rzeczy skupia się na możliwości odzwierciedlenia wszystkich elementów świata rzeczywistego w Internecie, ale także na wprowadzeniu „inteligencji" w przedmioty i umożliwieniu komunikacji pomiędzy nimi [21]. Aplikacje Internetu Rzeczy mogą być używane w wielu obszarach, takich jak: ochrona zdrowia, rolnictwo, inteligentne budynki (szkoły, szpitale, domy), zarządzanie łańcuchem dostaw, transport i obronność [22]. Po sukcesie mediów społecznościowych (które zyskiwały popularność począwszy od roku 2005) zespoły badawcze zaczęły podejmować próby połączenia idei Internetu Rzeczy i mechanizmów znanych z mediów społecznościowych [23]. W ten sposób powstała koncepcja Społecznościowej Sieci Rzeczy (ang. „Social Web of Things") [24], czy Społecznościowego Internetu Rzeczy (ang. „Social Internet of Things" – SIoT) [25]. Poziom ufności do innych urządzeń obecnych w sieci może zostać oszacowany na bazie oceny interakcji (relacji społecznych) pomiędzy agentami. Ustanowienie relacji zaufania jest możliwe na podstawie informacji rozproszonej czerpanej od wielu urządzeń, swoistego wzajemnego uwierzytelniania się [25]. Przykłady wykorzystania systemów zarządzania zaufaniem i reputację na potrzeby Internetu Rzeczy są dostępne w literaturze [26]–[30].

2.7.5. Sieci P2P

W przypadku sieci peer-to-peer, na przykład służących do wymiany plików, są także wykorzystywane systemy zarządzania zaufaniem i reputacją. Systemy TRM w sieciach P2P (opisane m.in. w publikacjach: [31]–[34]) mają za zadanie zachęcanie użytkowników do udostępniania plików o dobrej jakości (udostępnianie niewielkiej liczby plików, lub plików

o słabej jakości, prowadzi do zaniżenia reputacji, co przekłada się na ograniczenie możliwości pobierania plików).

2.7.6. Sieci WSN

Bezprzewodowe sieci sensorowe (WSN – ang. „Wireless Sensor Networks") także wykorzystują systemy zarządzania zaufaniem i reputacją do przetwarzania informacji pochodzących od poszczególnych sensorów, szczególnie w sytuacji gdy informacje te nie są spójne. Systemy TRM określają sposób podjęcia decyzji w sytuacjach konfliktowych, ze względu na to że podjęcie niewłaściwej decyzji może doprowadzić do awarii całego kontrolowanego środowiska. Propozycje takich systemów można znaleźć w artykułach: [35]–[38].

2.7.7. Sieci mobilne i ad-hoc

Na potrzeby mobilnych sieci ad-hoc (MANET – ang. „Mobile Ad-hoc NETworks") systemy zarządzania zaufaniem i reputacją są wykorzystywane do wymuszenia partycypacji w sieci przez poszczególne węzły poprzez przekazywanie (routing) pakietów. Węzły w tej sieci próbują racjonalnie gospodarować własnymi zasobami (np. posiadaną energią lub mocą obliczeniową) w związku z czym byłyby skłonne do ograniczenia własnej roli w przekazywaniu informacji generowanych przez inne węzły, w przypadku gdy nie jest to informacja użyteczna bezpośrednio dla węzła przekazującego. Z tego powodu systemy TRM (takie jak zaproponowane w publikacjach: [39]–[42]) mogą zapewnić bardziej sprawiedliwe i efektywne działanie protokołów routingu.

2.7.8. Wymiana informacji dotyczących bezpieczeństwa

W ramach wymiany informacji o zagrożeniach bezpieczeństwa (np. informacji o wskaźnikach kompromitacji, czy innych typów informacji użytecznej) niektóre zespoły reagowania na incydenty komputerowe wdrożyły systemy zarządzania zaufaniem i reputacją. Systemy TRM, w tym przypadku, mają na celu ocenę rzetelności źródła informacji, bądź samej informacji. Ma to istotne znaczenie ze względu na fakt, że użycie nierzetelnej informacji w systemach zapobieganiu włamaniom (IPS – ang. „Intrusion Prevention Systems") może skutkować zablokowaniem użytecznej komunikacji, prowadząc do odmowy usług (DoS – ang. „Denial of Service"). Podobnie, w przypadku użycia takiej informacji w systemach IDS (ang.

„Intrusion Detection Systems") lub systemach SIEM (ang. „Security Information and Event Management"), może ona prowadzić do wygenerowania dużej liczby fałszywych alarmów, które mogą odciągnąć uwagę analityka lub operatora systemu od rzeczywistych zagrożeń. Z drugiej strony, brak wykorzystania informacji o zagrożeniach bezpieczeństwa z możliwie wielu źródeł ogranicza skuteczność funkcjonowania systemów bezpieczeństwa. Wobec tego, wykorzystanie systemów zarządzania zaufaniem i reputacją może przynieść znaczące korzyści.

Przykładem wykorzystania systemu zarządzania zaufaniem i reputacją jest także ocena rzetelności informacji o podatnościach i eksploitach w ramach bazy danych agregujących informacje z różnych źródeł, która była współtworzona przez autora rozprawy i opisana m.in. w publikacjach: [43], [44].

2.7.9. Podsumowanie zastosowań systemów TRM

Systemy TRM są stosowane w różnych typach aplikacji, systemów i sieci. W przypadku wielu zastosowań wykorzystywane systemy TRM są bardzo proste, bazują jedynie na średniej ocenie, obliczonej na podstawie opinii użytkowników (agentów) i nie starają się uwzględnić możliwych działań zmierzających do zmanipulowania systemu. W wielu przypadkach można mieć wątpliwości czy tak proste mechanizmy oceny można nazwać systemem zarządzania zaufaniem i reputacją. W innych zastosowaniach (w szczególności tam, gdzie agenta w systemie nie można utożsamiać z użytkownikiem danego środowiska, jak np. w przypadku sieci WSN, MANET, czy P2P) stosowane są bardziej skomplikowane systemy TRM, które starają się a priori przyjąć, że niektóre z dostarczanych rekomendacji, czy też ogólniej, działań agentów, będą miały za cel zmanipulowanie systemu. Szerokie spektrum zastosowań systemów TRM uzasadnia potrzebę stworzenia metod oceny ich wiarygodności.

2.8. ATAKI NA SYSTEMY TRM

Atak przeciwko systemowi TRM może zostać zdefiniowany w następujący sposób:

***Definicja 8*: Atak na system TRM** *to wszelkie działania podejmowane przez agenta lub grupę agentów, które mają na celu zaburzenie lub zmianę rezultatu obliczeń miar zaufania lub reputacji, lub zmianę rezultatu procesu podejmowania decyzji przez innych agentów w oparciu o obliczone miary zaufania lub reputacji.*

Przedmiotem zainteresowania tej rozprawy są ataki bezpośrednio wymierzone w system zarządzania zaufaniem i reputacją, zgodnie z zaprezentowaną definicją powyżej,

w szczególności polegające na manipulowaniu wydawanymi rekomendacjami. Przykładem takiego ataku jest atak wychwalania, mający na celu zawyżenie oceny (rekomendacji) dotyczącej innego agenta. Przeprowadzenie tego ataku nie ma wszakże praktycznego sensu, a wręcz nie może on być przeprowadzony przy braku funkcjonowania systemu TRM (jeżeli nie funkcjonuje system TRM, to nie ma także wymiany rekomendacji pomiędzy agentami, nie są więc możliwe próby manipulowania nią).

Ataki wymierzone bezpośrednio w mechanizmy komunikacji pomiędzy agentami, np. zakłócanie transmisji, czy podszywanie się pod innego agenta, pozostają poza zakresem rozprawy. Ataki te nie są skierowane bezpośrednio przeciwko systemowi TRM, aczkolwiek system TRM może być wykorzystany do zapobiegania im. Specyficznym typem ataku jest atak kreacji wielu tożsamości – jest on wymierzony wprost w system TRM, ale najbardziej efektywne sposoby zapobiegania temu atakowi są oparte na mechanizmach wykraczających poza same systemy TRM (jak np. kontrola dostępu do systemu, czy stosowanie metod kryptograficznych).

Szerszy opis dobrze znanych, zidentyfikowanych w literaturze ataków przeciwko systemom TRM został dokonany w załączniku 4, a dodatkowo, w ograniczonym zakresie w rozdziale 3 (w podrozdziale 3.3). W dalszej części podrozdziału przedstawiono natomiast najistotniejsze typy ataków i ich wzajemne relacje.

Agenty stosujące ataki będą w dalszej części rozprawy określane jako agenty nierzetelne. Możliwe jest wyróżnienie dwóch podstawowych typów agentów nierzetelnych:

- agenty samolubne – dążące do maksymalizacji własnej korzyści z funkcjonowania w środowisku, kosztem korzyści osiąganych przez inne agenty i efektywności środowiska;
- agenty złośliwe – dążące do minimalizacji efektywności środowiska, w którym działają lub minimalizacji korzyści konkretnych innych agentów, natomiast niestarające się maksymalizować własnych korzyści, w związku z czym mogą mieć na celu zaburzenie działania środowiska nawet poprzez poniesienie własnych dodatkowych kosztów.

2.8.1. Ogólna klasyfikacja ataków na systemy TRM

Istniejące ataki przeciwko systemom TRM mogą być sklasyfikowane w sposób przedstawiony w tabeli 1.

Tabela 1 Ogólna klasyfikacja ataków na systemy TRM

Cecha	Typ ataku	Opis	Przykład
Zmienność ataku w czasie	Atak dynamiczny	atak, w którym zachowanie atakującego jest zmienne w czasie (na podstawie określonych, znanych atakującemu reguł)	atak oscylacji zachowania
	Atak statyczny	atak, w którym zachowanie atakującego jest stałe w czasie	atak stały
Zmienność ataku w zależności od stanu środowiska	Atak adaptacyjny	atak, w którym zachowanie atakującego (zarówno w kontekście świadczenia usług jak i wydawania rekomendacji) zmienia się w zależności od warunków panujących w środowisku, lub w określonej części środowiska albo określonego parametru środowiska lub systemu TRM (np. bieżącego zaufania do atakującego)	atak niespójnego zachowania
	Atak prosty	atak nieadaptacyjny	atak wychwalania
Liczba atakujących agentów i ich współpraca	Atak pojedynczego agenta	atak dokonywany przez jednego agenta	atak oscylacji zachowania
	Atak wielu agentów	atak wykonywany przez wielu agentów jednocześnie	atak oscylacji zachowania wykonywany przez wielu agentów
	Atak kooperacyjny	atak wielu agentów, w którym występuje forma współpracy pomiędzy agentami (np. synchronizacja działań lub specjalizacja działań agentów – występowanie ról)	grupowy atak oscylacji zachowania
Atak na usługę lub rekomendację	Atak na świadczenie usługi	atak, w którym działanie atakującego polega na nierzetelnym świadczeniu usługi (np. przekazywania pakietów, dostarczania plików, oferowania dóbr, itd.)	atak stały
	Atak na wydawanie rekomendacji	atak, w którym działanie atakującego polega na nierzetelnym wydawaniu rekomendacji lub opinii dotyczących innych agentów	atak oczerniania
Niedostarczanie lub fałszowanie rekomendacji	Atak utrudniania pozyskania rekomendacji	atak, w którym agenty nie dostarczają rekomendacji lub opinii na temat innych agentów	odmowa dostarczenia rekomendacji
	Atak fałszowania rekomendacji	atak, w którym dostarczone przez agenty rekomendacje lub opinie są zmanipulowane	atak oczerniania
Wykorzystanie fałszowania tożsamości agenta	Atak z fałszowaniem tożsamości	atak, w którym agent fałszuje swoją tożsamość, np. przedstawia się jako inny agent, lub kreuje wiele fałszywych tożsamości	atak kreacji wielu tożsamości
	Atak bez fałszowania tożsamości	atak, w którym nierzetelny agent nie dokonuje prób manipulowania własną tożsamością	inne niż atak kreacji nowej i wielu tożsamości

Warto podkreślić, że nie wszystkie z powyższych typów ataków są rozłączne, w szczególności istnieją dynamiczne adaptacyjne ataki kooperacyjne (np. atak wyrocznia zaproponowany przez autora rozprawy w artykule [45]).

2.8.2. Piramida bólu w odniesieniu do ataków na systemy TRM

Analogicznie jak w podrozdziale 2.2, gdzie została przedstawiona piramida bólu dotycząca ataków na systemy i sieci oparte o protokół IP, może zostać zaprezentowana piramida bólu w odniesieniu do ataków na systemy TRM. Piramida ta została stworzona przez autora pracy, na podstawie analizy poszczególnych typów ataków, a także sposobów ich wykrywania i reakcji na nie w systemach TRM[12]. Im niższy poziom piramidy, tym z jednej strony atak jest prostszy do wykrycia i tym więcej systemów TRM jest w stanie go wykryć (lub mu skutecznie przeciwdziałać), a z drugiej strony tym częściej jest badana odporność systemów TRM na te ataki.

Rysunek 7 Piramida bólu w odniesieniu do ataków na systemy TRM

Najtrudniejsze do wykrycia i przeciwdziałania są ataki adaptacyjne oraz kooperacyjne, one także są zwykle najbardziej skomplikowane i wymagają wysublimowanych działań ze

[12] Jeżeli na określonym poziomie piramidy nie podano typu ataku ze względu na określoną cechę, to oznacza, że uwzględnia on wszystkie typy ataków w ramach tej cechy, o ile na niższych poziomach piramidy ta cecha została wymieniona lub o ile cecha ta nie jest istotna. Np. ataki z fałszowaniem tożsamości pojawiają się dopiero na piątym poziomie piramidy (licząc od dołu). Oznacza to, że poziomy 1-4 dotyczą jedynie ataków bez fałszowania tożsamości. Zarówno ataki utrudniania pozyskania rekomendacji jak i ataki fałszowania rekomendacji dotyczą ataków w kontekście wydawania rekomendacji, dlatego począwszy od poziomu 3 ten typ ataków jest uwzględniony.

strony atakujących. Mimo że nie wynika to wprost z piramidy bólu, warto podkreślić, że ataki kooperacyjne są zwykle skuteczniejsze niż ataki wielu agentów, które wszakże są skuteczniejsze niż ataki pojedynczego agenta.

2.9. WIARYGODNOŚĆ SYSTEMÓW TRM

Miary wiarygodności systemów TRM zostaną przedstawione po wprowadzeniu modelu środowiska, systemu TRM i ataku, jednak ramowo można określić, że wiarygodnością systemu TRM jest pewna miara odporności systemu na ataki wymierzone w ten system. Z tego względu przyjęto następujące definicje dotyczące wiarygodności:

Definicja 9: ***System TRM jest wiarygodny****, jeżeli nie istnieje taki znany atak na system TRM, który jest w stanie istotnie zaburzyć (zmienić) decyzje podejmowane przez agenty w oparciu o ten system TRM.*

Definicja 10*:* ***Wiarygodność*** *jest cechą systemów TRM, będącą miarą poziomu odporności na ataki.*

Niniejsza praca ma za zadanie przedstawić metodykę oceny wiarygodności systemów TRM, w tym celu koncentruje się na przedstawieniu działań zmierzających do oceny odporności tego typu systemów na ataki przeciwko nim. W praktyce, rozprawa skupia się przede wszystkim na kontekście działań złośliwych agentów, choć także dotyka problematyki oceny odporności na działania samolubne.

3. STAN WIEDZY W ZAKRESIE SYSTEMÓW TRM

Przegląd literatury został dokonany z uwzględnieniem następujących typów prac:

- propozycji nowych systemów, w tym publikacji badających ich odporność na niektóre ataki;
- prac systematyzujących wiedzę o systemach TRM, w tym:
  - właściwości zaufania, reputacji i systemów TRM,
  - przeglądów i taksonomii,
  - teoretycznych opisów systemów;
- prac dotyczących ataków na systemy TRM, takich jak:
  - przeglądy istniejących ataków i propozycje nowych ataków,
  - taksonomie ataków;
- prac dotyczących sposobów oceny odporności systemów na ataki, w szczególności:
  - metod dokonywania oceny odporności,
  - propozycji miar ocen odporności,
  - propozycji narzędzi symulacyjnych do badań systemów TRM.

Zgodnie z powyższym podziałem, w kolejnych podrozdziałach zostanie przedstawiony stan wiedzy. Niektóre z publikacji dotyczą kilku wymienionych powyżej aspektów. W takich przypadkach ich krótkie omówienie może znajdować się w kilku podrozdziałach. Bieżący rozdział, mimo tego, że jest daleki od wyczerpującego przeglądu wszystkich prac związanych z tematyką systemów zarządzania zaufaniem i reputacją, to ma w założeniu zaprezentować i usystematyzować najczęściej pojawiające się propozycje i ocenić ich adekwatność oraz wskazać najistotniejsze ograniczenia w kontekście oceny wiarygodności systemów TRM.

3.1. PROPOZYCJE SYSTEMÓW TRM

Podrozdział zawiera przegląd propozycji systemów TRM. Ze względu na znaczną liczbę tego typu publikacji, wskazano jedynie wybrane z nich, stanowiące przykłady ich różnych zastosowań. Autor rozprawy zdecydował się na nieprzedstawianie opisów i specyfiki poszczególnych systemów, a jedynie ich wylistowanie i wskazanie najważniejszych wspólnych właściwości.

Ze względu na to, że zarządzanie zaufaniem i reputacją jest wymieniane często jako mechanizm zwiększenia bezpieczeństwa w sieciach WSN [46], istnieje wiele propozycji systemów TRM dla tego typu środowisk, jak na przykład EBTRM (Enhanced Bio-inspired Trust and Reputation Model), opisany w pracy [35], RFSN (Reputation-based Framework for

Sensor Networks) [47], czy HRFSN (Heider-theory based Reputation Framework for Sensor Networks) [36], jak również Lightweight Trust System with Provisioning for Detecting Malicious Node in Clustered Wireless Sensor Networks [37] oraz AraTRM (Attack resistible ant-based Trust and Reputation Model) [38]. Systemy TRM dla sieci WSN często skupiają się na zapewnieniu efektywnego przekazywania informacji w ramach sieci, nierzadko opierając swoje działanie na zapewnieniu skutecznego działania protokołów routingu. Większość z propozycji systemów TRM dla sieci WSN wykorzystuje możliwość obserwacji, czy inne węzły przekazują pakiety zgodnie z oczekiwaniami. Istotnym aspektem działania tych systemów są kwestie związane z zarządzaniem energią, z uwagi na fakt, że sensory w sieci WSN często mają ograniczone możliwości zasilania.

Dla sieci MANET stworzono systemy TRM, takie jak: TEREC (Trust Evaluation and Reputation Exchange for Cooperative intrusion detection in MANETs) [39], Reputation Based Trust Management System for MANET [40], CORE (A Collaborative Reputation Mechanism to enforce node co-operation in Mobile Ad hoc Networks) i CONFIDANT (Cooperation Of Nodes - Fairness In Dynamic Ad-hoc NeTworks) [47], czy system TRM zaproponowany w artykule [41]. W szczególności dla taktycznych sieci doraźnych zaproponowano system GlobalTrust [42]. Podobnie jak w przypadku sieci WSN, systemy TRM dedykowane dla sieci MANET koncentrują się na ocenie zaufania do węzłów w odniesieniu do przekazywania informacji, jednakże dodatkowym aspektem w ich przypadku są ciągłe zmiany topologii sieci.

Systemy TRM dla platform e-handlu najczęściej bazują na ocenach zaufania użytkowników w odniesieniu do innych użytkowników lub produktów i starają się wspomagać decyzję dotyczącą nawiązania interakcji z użytkownikami o największej rzetelności lub nabycia produktów o najwyższej jakości. Przykładami takich systemów są: ITRM (Iterative Trust and Reputation Mechanism) [14], [15], REF (Reputation Estimation Framework) [16], Fuzzy Logic Based Reputation Model [17], oraz system TATA (Temporal And Trust Analysis) [18].

Przykładami systemów TRM dla sieci P2P są: GenTrust (Genetic Trust management model for peer-to-peer systems) [31], SORT: A self-organizing trust model for peer-to-peer systems [32]. W oparciu o wcześniej zaprezentowane systemy TRM, tworzone są propozycje ulepszeń, czego przykładem jest system EigenTrust++ [33] utworzony na bazie systemu EigenTrust [34]. Systemy TRM dla sieci P2P starają się zapewnić uczciwą partycypację poszczególnych uczestników w środowisku, premiując udostępnienie większej liczby, dobrej jakości plików.

Dla innych zastosowań, m.in. do zarządzania wymianą multimediów lub informacji zaproponowano systemy TRM, takie jak np. system opisany w artykule [48], system ARTSense (Anonymity, Reputation and Trust in mobile sensing) [49], czy RLM [50].

Już ten bardzo skrócony przegląd propozycji systemów TRM ujawnia ich mnogość. Co więcej, systemy TRM są także tworzone do zupełnie niestandardowych zastosowań. Na przykład, autorzy artykułu [51] stworzyli pewien specyficzny system zarządzania zaufaniem i reputacją do oceny bezpieczeństwa oprogramowania. Zastosowana miara reputacji została oparta na informacjach o podatnościach konkretnego oprogramowania, w szczególności na szybkości eliminowania podatności, ale także zmian liczby ujawnianych podatności w jednostce czasu (równej jednemu miesiącu). Zaprezentowany system TRM wykorzystuje pojęcia zaufania i reputacji, ale w nieco inny sposób niż w większości propozycji systemów TRM: w systemie tym nie istnieje pojęcie rekomendacji oraz wymiany informacji o ocenach zaufania. Ewaluacja reputacji odbywa się jedynie na podstawie danych historycznych i pewnego rodzaju estymowanych oczekiwań w stosunku do przyszłości, również wykształconych na bazie danych historycznych. Z powyższych względów stosowanie przez autorów, w odniesieniu do własnej propozycji, pojęcia systemu zarządzania zaufaniem wydaje się nie do końca uprawnione. Z drugiej jednak strony jest to system, który wykorzystuje bezspornie w sposób prawidłowy pojęcie reputacji. Nie są rozważane ataki przeciwko temu systemowi. Propozycja ta wskazuje, że pojęcia zaufania i reputacji mogą być stosowane także dla tego typu celów.

Warto także podkreślić, że zwykle autorzy poszczególnych propozycji systemów TRM określają jedynie w jaki sposób jest wyliczane zaufanie lub reputacja, a pomijane są różne istotne zagadnienia w kontekście oceny wiarygodności, takie jak na przykład:

- jak często rekomendacje są przekazywane,
- jak następuje żądanie opinii,
- czy rekomendacje (opinie) są publiczne (dostępne także dla innych agentów) czy prywatne,
- w jaki sposób jest dokonywane uwierzytelnienie,
- czy wysyłane opinie są porównywane.

Brak określenia pełnej charakterystyki systemu TRM stwarza problemy z poprawną implementacją, a w konsekwencji może zwiększać podatność na ataki.

3.2. PRACE SYSTEMATYZUJĄCE WIEDZĘ O SYSTEMACH TRM

Podrozdział zawiera przegląd publikacji systematyzujących wiedzę o systemach TRM.

3.2.1. Prace dotyczące właściwości zaufania, reputacji i systemów TRM

Monografia [8] dotyczy zagadnień związanych z informatyką społeczną (ang. „social informatics"), przy czym definiuje to pojęcie jako „dyscyplinę informatyki badającą jak systemy informacyjne mogą realizować cele społeczne, używać pojęć z nauk społecznych lub stawać się źródłami informacji o zjawiskach społecznych" (ang. „*a discipline of informatics that studies how information systems can realize social goals, use social concepts, or become sources of information about social phenomena"*). Jako przykłady zastosowań pojęć zaufania (ang. „trust") i uczciwości (ang. „fairness") wymieniono w pracy: protokoły sieciowe, sieci P2P i ad-hoc, e-handel, wirtualne organizacje i przetwarzanie sieciowe (ang. „grid") oraz uczciwy podział zasobów w sieciach telekomunikacyjnych. W monografii przedstawiono definicje szeregu pojęć dotyczących zarządzania zaufaniem i uczciwością, takich jak: agent, interakcja (ang. „encounter"), akcja, kontekst, zarządzanie zaufaniem, zaufanie, reputacja, spolegliwość (ang. „trustworthiness"), ryzyko i niepewność, uczciwość, słuszność (ang. „equity"), sprawiedliwość (ang. „justice"), wzajemność (ang. „reciprocity") i altruizm. Pojęcia te zdefiniowano w oparciu o różne dyscypliny naukowe, nie tylko w oparciu o informatykę. Wiele z pojęć zdefiniowanych w publikacji [8] zostało wykorzystanych przez autora niniejszej pracy. Wymienione pojęcia zostały szeroko omówione w przytaczanej pozycji literaturowej, przy czym szczególnie dużo uwagi poświęcono teorii zaufania oraz teorii rozproszonej uczciwości.

Monografia [8] jest najbardziej obszerną i kompleksową, znaną autorowi niniejszej rozprawy, pozycją dotyczącą zagadnień związanych z mechanizmami zarządzania zaufaniem i reputacją i stanowiła jedną z głównych inspiracji do stworzenia niniejszej rozprawy. Nie porusza ona jednak szczegółowo zagadnień związanych z atakami na systemy wykorzystujące zarządzanie zaufaniem i reputacją, a w szczególności związanych z oceną odporności takich systemów na ataki, mimo tego, że dostrzega taką potrzebę i wskazuje istotne obserwacje dotyczące tego problemu. Publikacja w znacznej części dotyczy problemu uczciwości, np. uczciwego podziału dostępnych zasobów pomiędzy agentami, a nie bezpośrednio ataków na zaufanie i efektywność środowiska, co jest przedmiotem niniejszej pracy.

Zastosowania zarządzania zaufaniem i reputacją, zdaniem autora monografii [8], dotyczą otwartych, rozproszonych systemów (ang. „Open Distributed Systems" – ODS). Zgodnie z przedstawioną definicją, agent podejmuje decyzję nie tyle, z usług którego

z potencjalnych usługodawców skorzystać, ale którą z akcji (mogących być utożsamianymi z usługami) wykonać (jakiej usługi żądać). Celem zarządzania zaufaniem jest natomiast wsparcie agenta w podjęciu decyzji w warunkach niepewności, poprzez ustanowienie relacji zaufania lub nieufności z innymi agentami, od których akcji zależą rezultaty osiągnięte przez agenta podejmującego decyzję. W omawianej pracy nie zostało zdefiniowane pojęcie usługi, używane w niniejszej rozprawie, zamiast tego jest używane szersze pojęcie – akcji. Zaufanie zostało zdefiniowane jako relacja pomiędzy agentem ufającym (ang. „trustor") a agentem obdarzonym zaufaniem (ang. „trustee") w pewnym kontekście. Stosowane jest rozróżnienie pomiędzy zaufaniem ludzkim (ang. „human trust") i zaufaniem obliczeniowym (ang. „computational trust"). Zaufanie ludzkie jest utożsamiane z mentalnym stanem człowieka i jest przedmiotem zainteresowania psychologii, socjologii, antropologii, ekonomii i innych nauk. Zaufanie obliczeniowe jest reprezentacją koncepcji zaufania używaną w systemach zarządzania zaufaniem. W wielu przypadkach, systemy zarządzania zaufaniem starają się odzwierciedlić sposób podejmowania decyzji i wyciągania wniosków analogicznie jak w przypadku ludzkiego zaufania, mimo tego, że, ze względu na ogromne skomplikowanie odczuwania zaufania przez ludzi, może to być dokonane tylko w ograniczonym zakresie. W tym kontekście niniejsza rozprawa dotyczy tylko zaufania obliczeniowego. Pojęcie zaufania obliczeniowego jest także stosowane w innych pracach (np. [52]–[54]). Co istotne, pojęcie to nie zawsze jest używane w sposób jawny – prawie wszystkie pozycje bibliograficzne, przytoczone w tej rozprawie, dotyczą zagadnień zaufania obliczeniowego, mimo tego, że większość z nich nie używa przymiotnika „obliczeniowy" określając zaufanie. Zaufanie obliczeniowe ma możliwie dokładnie modelować zaufanie ludzkie. W efekcie badania dotyczące zaufania ludzkiego cechują się dużą istotnością w odniesieniu do zarządzania zaufaniem w ramach technik informacyjnych, w szczególności z uwagi na możliwość empirycznej weryfikacji.

Monografia [8] przedstawia także zestawienie definicji zaufania spotykanych w literaturze, zdecydowana większość przytoczonych definicji odnosi się jednak do aspektu socjologicznego lub psychologicznego zaufania i definiuje zaufanie ludzkie, a nie obliczeniowe, tym samym jest mniej interesująca w kontekście tej rozprawy i dlatego nie będzie przytaczana.

W pracy [8] spolegliwość (ang. „trustworthiness") została określona jako obiektywna, uzależniona od kontekstu jakość agenta w odniesieniu do zasługiwania na zaufanie. Jest to właściwość danego agenta, w przeciwieństwie do zaufania, które jest właściwością relacji pomiędzy grupą (typowo parą) agentów. Zdaniem autora [8] zdecydowanie łatwiej jest ocenić lub oszacować zaufanie niż spolegliwość. Reputacja jest informacją o agencie zaufanym, która

jest dostępna dla agentów ufających i jest wynikiem historii zachowania agenta zaufanego w odniesieniu do określonego kontekstu. Jest to zgodne z podejściem zaprezentowanym w niniejszej pracy, gdzie reputacja jest pewnego rodzaju właściwością agenta, zaufanie jest natomiast właściwością relacji pomiędzy parą agentów. Reputacja jest opartą na obserwowalnych objawach estymacją nieobserwowalnej z natury spolegliwości, obarczoną zawsze pewnym błędem (zwłaszcza wobec potencjalnej niestacjonarności spolegliwości). Zaufanie, podobnie jak reputacja, jest także estymacją spolegliwości, z tym, że dokonywaną indywidualnie, a nie systemowo. W omawianej monografii ryzyko jest definiowane jako oczekiwane wahania wyniku interakcji agenta, natomiast niepewność jako nieznane wahania wyniku interakcji. Uczciwość określono jako satysfakcję z usprawiedliwionych oczekiwań agentów uczestniczących w systemie, według reguł zaaplikowanych do określonego kontekstu w oparciu o motywy i przyczyny. Jakkolwiek zagadnienie uczciwości, zarówno poszczególnych agentów uczestniczących w systemie (ang. „individual fairness"), jak i całego systemu (ang. „system fairness"), jest niezmiernie istotne, o tyle w niniejszej rozprawie będzie mu poświęcona mniejsza uwaga. Rozpatrywanie zagadnienia uczciwości ma bowiem sens przy założeniu agentów mogących działać samolubnie, ale nie złośliwie. Niniejsza praca jest poświęcona przede wszystkim złośliwym atakom przeciw systemom TRM.

Książka [8] wskazuje także różne typy zaufania: zaufanie dotyczące oczekiwań (ang. „expectancy trust"), zaufanie dotyczące zależności (ang. dependency trust), zaufanie afektywne (ang. „affective"), kognitywne (ang. „cognitive") i nieufność (ang. „distrust"), jak również zaufanie dotyczące rzetelności (ang. „credibility trust"). Te typy zaufania mają znaczenie w odniesieniu do ludzkiego zaufania, a nie bezpośrednio do zaufania obliczeniowego. Wskazywane są również typy reputacji (indywidualna, grupowa, bezpośrednia, pośrednia, itd.) oraz różnice pomiędzy pojawiającymi się w literaturze pojęciami lokalnej i globalnej reputacji. Globalna reputacja jest związana ze spolegliwością agenta, a nie z zaufaniem.

Monografia [8] ujmuje także zagadnienia zarządzania zaufaniem i reputacją w kategoriach teorii gier, przywołując, między innymi, szeroko znany problem dylematu więźnia [55], w szczególności w wersji iteracyjnej (powtarzanej wielokrotnie rozgrywki). Najprostszą i jedną z najbardziej efektywnych strategii w tej grze jest strategia „reputacyjny wet za wet", która polega na tym, że w pierwszej iteracji decyzja o tym czy współpracować, czy oszukiwać drugiego gracza jest uzależniona od jego reputacji (lub zaufania do tego gracza), a następnie stosowane jest zachowanie w kolejnej iteracji zgodne z zachowaniem przeciwnika z poprzedniej iteracji. Jest to pewna modyfikacja strategii „wybaczający wet za wet", polegająca na uwzględnieniu reputacji drugiego gracza [8].

Praca [8] porusza również właściwości mechanizmu propagacji zaufania, między innymi takie jak: przechodniość (ang. „transitivity”), wzajemność (ang. „reflexivity”), czy podobieństwo (ang. „similarity”).

3.2.2. Przeglądy i taksonomie

Publikacja [56] stanowi próbę zebrania informacji i przeglądu różnych systemów TRM w celu dążenia do standaryzacji i stworzenia ogólnego modelu takich systemów. W pracy zostały krótko opisane następujące systemy TRM: Sporas, Regret, AFRAS, MTrust (przeznaczone dla szerokiego spektrum systemów wieloagentowych); DWTrust, AntRep, EigenTrust (przeznaczone dla sieci P2P); RRS, PTM (przeznaczone dla sieci ad-hoc); QDV, ATRM (przeznaczone dla sieci WSN). Przedstawiono także ogólną, krótką klasyfikację wymienionych systemów TRM, opierając się w głównej mierze na tym, czy wykorzystują pojęcia zaufania, czy reputacji oraz jakiej techniki używają do ewaluacji miary zaufania lub reputacji (spośród następujących technik: Fuzzy, Bayesian, Bio-inspired, Social-network, Analytic).

W artykule [57] przedstawiono przykłady systemów wykorzystujących pojęcia zaufania lub reputacji, m.in. takie jak: m-RACER, EMPIRE, K-FTM, TRANS, RFSN, ATRM (używane w celu zabezpieczenia protokołów routingu); TIBFIT, TrustVoting (wykorzystywane do wykrywania źle działających sensorów); czy protokół SELDA jako mechanizm służący do bezpiecznej agregacji danych.

Praca [58] zawiera przegląd stanu sztuki związanego z systemami TRM – w szczególności opis kilkunastu systemów TRM, takich jak: CuboidTrust, EigenTrust, BNBTM, GroupRep, AntRep, Semantic Web, Global Trust, PeerTrust, PATROL-F, Trust Evolution, TDTM, TACS. Systemy te zostały stworzone do stosowania w sieciach P2P. Omawiane źródło zawiera także porównanie sposobów działania TRM w oparciu o nierozbudowaną klasyfikację, a także podsumowanie dotyczące słabych i mocnych stron poszczególnych systemów. W kontekście ataków na systemy TRM wskazano, że ataki, oparte w szczególności na kooperacji wielu agentów, są możliwe i żaden z przedstawionych systemów nie jest całkowicie odporny na takie ataki. Jako przykład ataku został wymieniony atak kreacji wielu tożsamości. Podkreślono także potrzebę stworzenia narzędzi do walidacji systemów TRM, umożliwiających testowanie każdego mechanizmu zarządzania zaufaniem i reputacją. Opracowanie skupia się na sieci P2P, jednakże ten wniosek bez trudu można także rozszerzyć na inne typy środowisk działania systemów TRM.

W artykule [59] zawarto przegląd i klasyfikację wybranych systemów zarządzania zaufaniem i reputacją. Klasyfikacja została stworzona z uwzględnieniem aspektów takich jak: źródła informacji, widoczność ocen reputacji lub zaufania, granularność (możliwość uwzględniania różnych kontekstów) i typy wymienianych informacji.

Przegląd różnych systemów zaufania w podziale na różne środowiska ich wykorzystania, ze szczególnym uwzględnieniem sieci WSN, zawarto w dokumencie [60]. Przeprowadzono także szeroką dyskusję zastosowań systemów TRM w różnych środowiskach, takich jak e-handel, sieci P2P, sieci WSN i ad-hoc. Wymieniono również przykłady ataków przeciwko systemom TRM, takie jak atak kreacji wielu tożsamości, czy oczernianie, ale bez szerszej dyskusji. Autorzy dokonali ponadto klasyfikacji systemów TRM z uwzględnieniem przyjętego sposobu wyliczania zaufania, np. w oparciu o prawdopodobieństwo, modele Bayesa lub logikę rozmytą oraz z uwzględnieniem wykorzystywanych źródeł informacji (np. obserwacje bezpośrednie lub pośrednie).

Publikacja [61] zawiera skrócony opis sześciu systemów zarządzania zaufaniem i reputacją: FIRE, REGRET, system stworzony przez Yu i Singh [62], TRAVOS, PeerTrust oraz BRS, a także prezentuje propozycję klasyfikacji systemów TRM w odniesieniu do takich kryteriów jak: sposób liczenia zaufania (model probabilistyczny versus deterministyczny), źródła informacji, uwzględnienie kontekstu, adaptowalność, czy sposób ewaluacji rzetelności ocen. Jako jeden z otwartych problemów związanych z systemami TRM, autorzy wymieniają ocenę ich odporności na wysublimowane (nietypowe i rozbudowane) ataki.

Praca [63] zawiera przegląd systemów TRM dla sieci P2P, a także omówienie właściwości zaufania. Uwzględniono w niej przykłady szeroko znanych ataków przeciwko systemom TRM, a także przedstawiono propozycje obron. Autorzy wskazują, że znaczącym zagrożeniem dla systemów TRM są ataki kooperacyjne.

W monografii [8] wskazano pewne cechy systemów zarządzania zaufaniem i reputacją, które są pomocne przy ich klasyfikacji, takie jak np. typ używanego dowodu, zakres agregacji, obliczana wartość, czy odporność systemu. W szczególności ta ostatnia cecha mogłaby wydawać się interesująca, jednak nie dotyczy ona miary odporności systemu na ataki, a jedynie typu używanych mechanizmów dla zapewnienia tej odporności.

W artykule [64] został zaproponowany system TRM, który z wykorzystaniem algorytmów uczenia maszynowego, pozwala na wykrywanie agentów dostarczających usługi o niesatysfakcjonującej jakości lub dokonujących ataków. W publikacji tej dokonano ewaluacji zaproponowanego systemu w odniesieniu do wykrywania najpopularniejszych niekooperacyjnych ataków, stosując miary wykorzystujące liczbę właściwie i niewłaściwie

sklasyfikowanych agentów. Artykuł [64] zawiera także przegląd publikacji zawierających propozycje systemów TRM i określenie jakie ataki są rozważane przez ich autorów.

3.2.3. Teoretyczne opisy systemów TRM i modelowanie systemu TRM

Monografia [8] postuluje, że w przyszłości zarządzanie zaufaniem i reputacją może stać się nową, zdefiniowaną standardową usługą bezpieczeństwa, obok poufności, integralności, dostępności, uwierzytelnienia, autoryzacji, czy rozliczalności. Autor prezentuje też własny model funkcjonowania systemu zarządzania zaufaniem. Wskazuje w nim między innymi na obiekt „dowodu" (ang. „proof"), który może być dowolną informacją pozwalającą na obliczenie zaufania (np. rekomendacja, raport, lub obserwacja). W niniejszej rozprawie autor, prezentuje odmienny model, ale w wielu aspektach zbieżny z modelem przedstawionym w omawianym źródle.

Interesującą publikacją, w której podjęto próbę stworzenia modelu systemu zarządzania zaufaniem i reputacją, mającego znaleźć zastosowanie do ewaluacji odporności na ataki, jest artykuł [65]. Warto także zaznaczyć, że artykuł ten cytuje jedną z wcześniejszych prac autora tej rozprawy, służącą przedstawieniu koncepcji przydatnych do stworzenia modelu umożliwiającego ocenę odporności systemów zarządzania zaufaniem i reputacją [66]. Omawiane źródło porównuje własny model ze wstępnym modelem zaprezentowanym przez autora rozprawy i wykazuje pewne przewagi własnego modelu. Wskazane ograniczenia modelu autora rozprawy zostały dostrzeżone i zawarte explicite w samym artykule, który opisywał jedynie wstępny etap tworzenia kompleksowego modelu systemu i ataku. Praca [65] została opublikowana w czasie tworzenia niniejszej rozprawy, prezentując jednak odmienne podejście do stworzenia modelu systemu TRM służącego ewaluacji. Warto podkreślić, że o ile głównym celem autorów artykułu [65] było stworzenie modelu systemu TRM, który może być wykorzystany do ewaluacji takich systemów (co zresztą zostało dokonane na kilku przykładach w ramach artykułu), to głównym celem tej rozprawy jest opracowanie metodyki ewaluacji oraz dedykowanego narzędzia służącego do tego celu, a także przeprowadzenie na tej podstawie rzeczywistych badań opartych na własnym modelu systemów TRM. Tak więc niniejsza praca w pewnym sensie wykracza poza przedmiot zainteresowania cytowanego artykułu. Zarówno model systemu TRM przedstawiony w tej rozprawie, jak i model opracowany przez autorów artykułu [65], posiadają wiele cech wspólnych, ale są też w znacznym stopniu odmienne. Mimo braku całkowitej zbieżności, wydaje się, że nie są ze sobą sprzeczne, a jedynie koncentrują się

na nieco innych aspektach. Zdaniem autora rozprawy, zaprezentowany przez niego model może być w prostszy sposób wykorzystany przy rzeczywistych badaniach systemów.

Publikacja [65] wskazuje, że większość prac dotyczących analizy bezpieczeństwa systemów TRM nie jest oparta na odpowiednim modelu formalnym, a jedynie dokonuje intuicyjnych ewaluacji w odniesieniu do szczególnych przypadków. Zdaniem autorów istnieje także wiele prac które dokonują ewaluacji systemów TRM za pomocą metod symulacyjnych, jednakże ich istotną charakterystyką jest brak ogólności oraz ograniczona stosowalność. Wnioski te są zbieżne z obserwacjami autora niniejszej rozprawy. Autorzy omawianej publikacji twierdzą także, że zaprezentowany przez nich model systemu TRM – TRIM (Trust and Reputation Interaction Model) jest w stanie objąć szerokie spektrum systemów TRM, oraz jest możliwe na jego podstawie zdefiniowanie zaawansowanych i skomplikowanych ataków. W celu zaprezentowania stosowalności modelu, kilka znanych systemów TRM, łącznie z wybranymi istotnymi atakami, zostało zdefiniowanych w oparciu o zaprezentowany model. Dodatkowo, zostało stworzone narzędzie TRIM-checker służące do wstępnej weryfikacji odporności systemów TRM na ataki. Jak podkreślają jednak sami autorzy, narzędzie to posiada wciąż znaczące ograniczenia. W pracy tej podkreślany także jest fakt, że mimo tego, że różne systemy TRM mają różną odporność na ataki, wciąż nie ma kompleksowej metody ewaluacji odporności tych systemów na ataki, a pierwszym krokiem do tego celu jest stworzenie modelu systemu TRM funkcjonującego we wrogim środowisku. Z całą pewnością jest to bardzo ważna praca w kontekście stworzenia modelu systemu TRM, służącego do ewaluacji ich odporności na ataki, ale wciąż wydaje się być niewystarczająca pod kątem dokonania praktycznej ewaluacji systemów TRM. Przede wszystkim, autorzy nie biorą pod uwagę niektórych właściwości systemów TRM, które są zależne od środowiska w jakim funkcjonują. Mimo twierdzenia autorów [65] o stworzeniu kompleksowego modelu, przyjęto szereg założeń upraszczających, także nie wyrażonych explicite. Przykłady takich założeń i ograniczeń są następujące:

- w modelu nie są brane pod uwagę opóźnienia możliwości dokonania oceny względem momentu świadczenia usługi, tak więc niektóre typy ataków nie mogą być uwzględnione;
- autorzy posługują się przykładami środowiska złożonego maksymalnie z 4 agentów, co więcej, jak sami zaznaczają, w środowisku złożonym z większej liczby agentów ewaluacja jest uniemożliwiona wskutek znacznej złożoności obliczeniowej;
- atak przeprowadzany przez złośliwe agenty musi być określony a priori (nie jest możliwe dostosowywanie strategii atakujących do rzeczywistego stanu środowiska),

czyli wszelkie ataki adaptacyjne, stanowiące największe zagrożenie dla systemów TRM, nie mogą być uwzględnione;

- analiza może być przeprowadzona jedynie dla z góry ustalonych parametrów systemu TRM, nie jest możliwa optymalizacja doboru parametrów (problem ten wskazują sami autorzy);
- rozważana jest jedna, homogeniczna usługa, która może być świadczona w środowisku; nie jest możliwe ujęcie w modelu większej liczby usług;
- autorzy nie rozważają ataku Sybil (kreacji wielu tożsamości), ani o nim nie wspominają;
- nie ma możliwości modelowania różnych klas zaufania lub reputacji (np. zaufania rekomendacyjnego – do wydawanych rekomendacji, czy zaufania akcyjnego – do jakości świadczonej usługi);
- rozważane są jedynie skrajne wartości poszczególnych parametrów czy miar zaufania (np. skrajne wartości rekomendacji), tj. maksimum i minimum, podczas gdy w praktyce możliwe jest stosowanie (np. przez atakujących) całego spektrum możliwych wartości, co może mieć znaczący wpływ na efektywność ataku.

Warto także podkreślić, że za pomocą swojego narzędzia i modelu autorzy przeprowadzili kalkulacje jedynie z uwzględnieniem pierwszych 10 interakcji. Co więcej, sami autorzy zauważają, że po około 20 interakcjach dochodzi do „eksplozji przestrzeni stanów", co w praktyce uniemożliwia dalszą ewaluację. W związku z tym wszelkie ataki, które polegają na zbudowaniu przez atakujących wysokiego poziomu zaufania/reputacji, a dopiero później zakładają przystąpienie do przeprowadzania rzeczywistych ataków, nie mogą być zamodelowane.

Praca [65] wskazuje dwa typy środowisk: peer-to-peer oraz klient-serwer. Różnica pomiędzy nimi polega na tym, że w pierwszym przypadku agent może w pewnych interakcjach pełnić funkcję klienta, a w innych serwera, podczas gdy w drugim typie środowiska role są ustalone. Ograniczeniem publikacji wydaje się być brak uwzględnienia zastosowań systemów TRM np. w problemach routingu, gdzie podczas jednej interakcji, dla przekazania pakietu, które się zakończy sukcesem, może być konieczny udział wielu usługodawców. Dodatkowo, w tym przypadku, klient nie zawsze ma możliwość dokonania oceny indywidualnej poszczególnych usługodawców, którzy uczestniczyli w przekazywaniu pakietów na danej trasie.

W stosunku do przytaczanej pracy [65] jedynym aspektem, który nie został ujęty w modelu zaprezentowanym w tej rozprawie, jest możliwość odmowy świadczenia usługi

przez usługodawcę. Sytuacja, kiedy jest to możliwe, jest jednak rzadka w rzeczywistych środowiskach i dodatkowo może być utożsamiana z pewnego rodzaju dyskryminacją. Co więcej, rozszerzenie modelu o taki przypadek nie jest szczególnie trudne, ale prowadzi do, jak się wydaje, niepotrzebnego wzrostu jego skomplikowania.

Warto także zaznaczyć, że ocena prac innych autorów, dokonana w artykule [65], jest przeprowadzona na zasadzie uznaniowej i nie zawsze właściwie odzwierciedla rzeczywistość. Dla przykładu, artykuł autora rozprawy [66] jest postrzegany jako rozpatrujący jedynie rozproszone systemy reputacyjne, a kompletnie pomijający aspekt scentralizowanych i rozproszonych systemów zaufania oraz scentralizowanego systemu reputacyjnego, podczas gdy artykuł ten wskazuje wszystkie cztery kategorie systemów wedle definicji przyjętej przez autorów pracy [65]. Także inne aspekty dokonanego porównania wydają się mieć na celu wsparcie tezy o przewadze własnego modelu zaprezentowanego przez autorów. Mimo powyższych uwag należy podkreślić, że praca stanowi ważny krok na drodze do ilościowego porównywania systemów zarządzania zaufaniem i reputacją.

Publikacja [56] podejmuje próbę określenia wymagań dotyczących zawartości opisu systemów TRM w celu dążenia do standaryzacji i poniekąd stworzenia ogólnego modelu takich systemów. Wskazuje ona że funkcjonowanie systemów TRM można podzielić na pięć głównych faz: zbieranie informacji, ocena i tworzenie rankingu, wybór agenta, transakcja oraz nagradzanie i karanie. Taki podział znajduje także odzwierciedlenie w innych pracach, m.in. w [67]. Model systemu TRM został także zaproponowany w publikacji [68], jednak jego użycie do ewaluacji odporności systemów TRM na ataki wydaje się nie być przydatne.

3.3. PRACE DOTYCZĄCE ATAKÓW NA SYSTEMY TRM

W niniejszym podrozdziale dokonano przeglądu literatury pod kątem opisu ataków przeciwko systemom TRM, przy czym w większości przypadków pominięto opis przebiegu czy właściwości ataku. Zestawienie najistotniejszych zidentyfikowanych ataków, łącznie z ich krótką charakterystyką, zostało zawarte w załączniku 4.

3.3.1. Opisy ataków

W wielu publikacjach (np. [1], [9], [69]) przytaczane są opisy i charakterystyki najprostszych ataków, takich jak *atak stały* (ang. „constant”), atak *oscylacji zachowania* (ang. „on-off”), a także ataków polegających na manipulowaniu dostarczanymi rekomendacjami

poprzez zawyżanie (atak *wychwalania* – ang. „false-praise") czy zaniżanie (atak *oczerniania* – ang. „bad-mouthing") ocen innych agentów. Często (np. w artykułach [1], [9], [69]) przytaczane są także opisy i analizy ataków wymierzonych w identyfikację agentów, w szczególności atak *kreacji wielu tożsamości* (ang. „Sybil attack"), czy atak *kreacji nowej tożsamości* (ang. „Whitewashing").

Warto również zaznaczyć, że niektóre prace używają nazw ataków w różnych znaczeniach. Artykuł [70] wymienia atak *Constant* – jako polegający na wydawaniu nieprawidłowych rekomendacji o innych agentach (wysokie rekomendacje o agentach złośliwych, a niskie o agentach rzetelnych). Jest więc to kombinacja ataków *wychwalania* i *oczerniania*. Tymczasem, nazwa ataku *Constant* w zdecydowanej większości pozycji literaturowych odnosi się do ataku polegającego na wykonywaniu w sposób ciągły nierzetelnych akcji (świadczeniu usług w sposób nierzetelny), a nie do manipulowania rekomendacją. Co więcej, niejednokrotnie ten sam atak bywa różnie nazywany przez różnych autorów.

Publikacja [70], oprócz opisanego powyżej ataku *Constant,* opisuje także pięć innych ataków. Atak *Camouflage* – jest pewną kombinacją ataku *Constant*, w którym atakujący agent wydaje generalnie złe rekomendacje, podobnie jak w powyższym ataku, ale strategicznie (kiedy przyniesie mu to znaczną korzyść) wydaje też rekomendacje prawidłowe (np. w celu zbudowania zaufania do siebie). W artykule jednak brak jest określenia, na podstawie jakich kryteriów i w jakim momencie złośliwy agent podejmuje decyzję o stosowaniu manipulacji rekomendacjami. Kolejny ze zidentyfikowanych przez autorów ataków – *Whitewashing* – jest analogiczny do przytoczonego wcześniej w tym rozdziale ataku polegającego na kreacji nowej tożsamości. Kolejne trzy ataki prezentowane w przywoływanym artykule – *Sybil Constant, Sybil Camouflage,* i *Sybil Whitewashing* są analogiczne odpowiednio do ataków *Constant, Camouflage* i *Whitewashing*, z tą tylko różnicą, że w ich przypadku zakłada się, że atakujących agentów jest więcej niż połowa spośród wszystkich agentów w danym środowisku. Przewaga liczebna atakujących może zostać osiągnięta przez rzeczywistą obecność wielu agentów atakujących, albo dzięki wykorzystaniu *ataku kreacji wielu tożsamości*. Warto zwrócić uwagę, że ataki prezentowane w omawianym źródle są pewnego rodzaju kombinacją ataków najczęściej przytaczanych w literaturze, ale używane nazewnictwo dla tych ataków wprowadza istotne niejednoznaczności w odniesieniu do przeważnie stosowanej nomenklatury.

W kontekście ataków przeciwko systemom zarządzania zaufaniem i reputacją, książka [8] wskazuje jako przykład atak *oszustwa po raz pierwszy* (ang. „first-time cheating"), który objawia się nagłą zmianą w charakterystyce zachowania agenta. Wskazuje także na możliwość

tworzenia koalicji przez atakujących w celu zwiększenia wpływu nieprawidłowych rekomendacji (nazywanych raportami). Jako przykład ataku na reputację jest także wymieniana *dyskryminacja* (ang. „discrimination"), będąca odpowiednikiem ataku *niespójnego zachowania*, wymienianego w innych publikacjach.

Przytaczając fragment artykułu [71] autora tej rozprawy warto zauważyć, że *„niektóre pozycje literaturowe* [69], [72] *podkreślają fakt możliwości istnienia kooperacyjnych ataków na systemy zarządzania zaufaniem i konieczność dokładnego zbadania tego typu strategii, ale nie podają one stosownych przykładów, ani dokładnej charakterystyki (elementów specyficznych) tych ataków. Zdaniem autorów pracy* [72], *inteligentne ataki w większości będą oparte na kooperacji wielu złośliwych agentów (kliki) w celu wzajemnego zbudowania dobrej reputacji tych agentów, a następnie jej wykorzystania do stworzenia bezpośredniego zagrożenia dla funkcjonowania środowiska, albo w celu zdyskredytowania określonych, prawidłowo funkcjonujących agentów. Podkreślana jest także konieczność stworzenia odpowiednich metod wykrywania takich ataków na bazie identyfikacji ponadprzeciętnie silnych powiązań pomiędzy niektórymi agentami. Jednocześnie omawiane źródło podkreśla, że w literaturze zbyt mało uwagi poświęcono kooperacyjnym atakom, tzn. takim, w których uczestniczy wiele atakujących agentów w sposób skoordynowany."*

Artykuł [71] przywołuje także inny przykład ataku zidentyfikowanego w literaturze: *„w artykule* [73] *rozważany jest specyficzny przypadek zastosowania systemu zarządzania zaufaniem: agenty określają zaufanie tylko do agenta sąsiedniego, a korzystają przy tym z rekomendacji dostarczonych także jedynie przez agenty sąsiednie (tzn. takie, z którymi są utrzymywane bezpośrednie relacje). Dla tego systemu została zaprezentowana propozycja ataku kooperacyjnego, a także sposobu obrony przed nim (poprzez wykrywanie konkretnego zachowania atakujących). Warty podkreślenia jest fakt, że zaprezentowany system zarządzania zaufaniem jest dość prostym i w związku z tym podatnym na ataki. Zarówno sam sposób ataku jak i obrony nie może zostać zastosowany w przypadku bardziej rozbudowanych systemów. (...)*

W artykule [74] *przedstawiono sześć strategii ataków opartych na kooperacji wielu agentów. Dla każdej z tych strategii zaproponowano sposób obrony, ale brak jest propozycji spójnego mechanizmu, który byłby w stanie wyeliminować wszystkie zaprezentowane ataki. Warto podkreślić, co czynią też autorzy przywoływanej publikacji, że ataki kooperacyjne są groźne dla większości (lub wszystkich) systemów zarządzania zaufaniem i reputacją. Co więcej, istnieje prawdopodobnie wiele innych ataków wykorzystujących kooperację, w przypadku których implementacja nawet wszystkich opracowanych do tej pory mechanizmów obronnych byłaby niewystarczająca.*

W artykułach [10], [75] *zaprezentowano grupowy atak oscylacyjny, który polega na tym, że w jednym czasie pewna grupa atakujących agentów dostarcza nieprawidłowe rekomendacje dla agentów będących celem ataku. Inna grupa atakujących stara się zwiększyć swoją reputację poprzez dostarczanie prawidłowych rekomendacji dla pewnych agentów, które nie są szczególnie istotne (znajdują się poza celem ataku). Po pewnym czasie, te dwie grupy atakujących agentów zamieniają się rolami. Atak ten jest nieskomplikowany, jak na atak kooperacyjny, dlatego wystarczającym sposobem obrony wydaje się być zaproponowany przez autorów mechanizm polegający na przywiązywaniu większej wagi do nierzetelnych zachowań agentów niż do zachowań prawidłowych.*

W ataku RepTrap opisanym w publikacjach [10], [76] *złośliwe agenty dokonują analizy interakcji zachodzących między rzetelnymi agentami i wybierają jako cel ataku te spośród rzetelnych agentów, które cieszą się wysoką reputacją, jednocześnie mając niewiele interakcji z innymi agentami. Wtedy też złośliwe agenty przeprowadzają atak oczerniania – dostarczając wszystkim agentom w środowisku negatywne rekomendacje na temat wybranych agentów."*

W artykule [45] autor niniejszej rozprawy zaprezentował atak kooperacyjny polegający na współpracy dwóch grup złośliwych agentów (nazwanych „wyroczniami" oraz „obserwowanymi"). Istotą tego ataku jest teoretycznie sprzeczne z ich interesami zachowanie złośliwych agentów – polegające na tym, że „wyrocznie" zawsze dostarczają prawidłowych rekomendacji na temat przyszłych zachowań agentów „obserwowanych". Takie działanie, w połączeniu z oscylacją zachowania agentów „obserwowanych", pozwala na uzyskanie przez złośliwe agenty wyższej reputacji niż przez agenty rzetelne. Autor zaprezentował skuteczne mechanizmy obrony przed tym atakiem, ale przykład tego ataku wskazuje na to, że może istnieć jeszcze wiele ataków bazujących na kooperacji, dla których opracowane do tej pory mechanizmy obronne będą nieskuteczne.

W pracy [57] znajduje się opis ataków, którym przeciwdziałają systemy zarządzania zaufaniem i reputacją, ale są także wskazane ataki wymierzone bezpośrednio w systemy TRM. Przykładem takiego ataku jest *atak inteligentnego zachowania* (ang. „intelligent behavior attack"), w którym adwersarz automatycznie dopasowuje własne zachowanie na podstawie przechwyconych informacji, co wpływa na ocenę zaufania lub reputacji. Adwersarz w takim ataku może selektywnie dostarczać usługi o złej lub dobrej jakości, a także zawyżać lub zaniżać rekomendacje dotyczące innych agentów. Warto zauważyć, że w praktyce nie jest to opis konkretnego ataku, a jedynie charakterystyka typu ataku, określonego w tej pracy jako ataki adaptacyjne. Omawiane źródło głównie koncentruje się na wykorzystaniu systemów TRM w sieciach WSN, z tego też względu podaje zestaw jedenastu wymagań, które powinny zostać

spełnione w celu zapewnienia bezpieczeństwa. Zaprezentowane wymagania są wszakże bardzo generyczne, obejmują m.in. standardowe usługi bezpieczeństwa np. uwierzytelnienie, autoryzację, zapewnienie poufności, czy niezaprzeczalności.

W publikacji [77] zaproponowano mechanizm detekcji złośliwych użytkowników (także współpracujących ze sobą) w ramach platform e-handlu. Mechanizm ten miał stanowić próbę ograniczenia negatywnego wpływu ataków. Zaproponowane podejście skupia się na detekcji anomalii (np. na bazie wykrywania ponadstandardowej różnorodności wystawianych ocen). W artykule dokonano także przeglądu prac, które starają się zidentyfikować złośliwych użytkowników, ale zdaniem autorów w większości przypadków nie są one skuteczne w wykrywaniu wczesnych faz ataków. Zostały także wykonane symulacje pokazujące, że używając zaproponowanego mechanizmu, złośliwi użytkownicy zostali właściwie wykryci. Omawiane źródło prezentuje interesujące podejście do wykrywania ataków (także kooperacyjnych), ale nie stanowi znaczącego przełomu w kwestii oceny wiarygodności systemów TRM.

Praca [78] porównuje odporność wybranych systemów przeznaczonych dla platform e-handlu na kilka strategii ataków. Autorzy zauważają, że badania symulacyjne nowych propozycji systemów zwykle ograniczają się do badania ich odporności na proste ataki pojedynczych agentów, którzy oszukują w sposób losowy. W praktyce jednak należy dążyć do badania systemów w kontekście odporności na działania wielu (potencjalnie współpracujących ze sobą) agentów, którzy działają „inteligentnie" (tzn. dążą do realizacji obranego przez siebie celu). Badania systemów zostały wykonane we własnym symulatorze i wykorzystano w nich następujące ataki:

- Proliferation (atak polega na kreacji wielu tożsamości);
- Countermeasures (w przypadku gdy istnieje wiele instancji nierzetelnego sprzedającego, szansa, że bez wcześniejszych informacji kupujący wybierze rzetelnego sprzedającego maleje, atak ten jest zdaniem autora rozprawy specyficznym przypadkiem wykorzystania i rozwinięciem ataku Proliferation);
- Reputation Lag (atakujący zachowują się rzetelnie przez 45 dni, a przez 15 kolejnych dni nierzetelnie, po tym okresie na nowo wchodzą do środowiska);
- Re-entry (atakujący zachowują się nierzetelnie, a po pewnym czasie ponownie wchodzą do środowiska z nową tożsamością);
- Value Imbalance (atakujący rzetelnie sprzedają tanie produkty, a oszukują na drogich, przy czym starają się utrzymać swoją reputację na określonym poziomie);

- Initial Window (atakujący wykorzystują początkowy okres obecności nowych agentów w środowisku, w którym nie mają oni wystarczającego doświadczenia wynikającego z własnych interakcji);
- Exit (atakujący opuszczają środowisko po wykonaniu wielu nierzetelnych interakcji).

Warto zauważyć, że autorzy często określają jako odmienny atak zasadniczo te same działania atakujących, podkreślając jedynie ich różne efekty. Rozważane jest także połączenie wszystkich wymienionych powyżej ataków. Autorzy zauważają, że – w zależności od systemu TRM – najbardziej skuteczne strategie atakujących mogą być różne (a atakujący mogą nie wiedzieć jaki system jest stosowany), ale dodatkowo twierdzą, że wcale nie jest to ograniczeniem dla atakujących, którzy mogą jednocześnie stosować wiele ataków i wybierać te, które są najbardziej skuteczne. Autorzy podkreślają, że strategie ataków nie zostały zoptymalizowane, tzn. parametry ataków nie zostały dobrane tak, aby zapewnić największą skuteczność ataku. Optymalizacja tych strategii została zidentyfikowana jako przyszły temat badań. Interesującymi zagadnieniami badawczymi zdaniem autorów będzie także zagadnienie wyboru ataku w przypadku, gdy nie jest możliwe stosowanie jednocześnie wielu ataków, oraz poddanie badaniom większej liczby systemów TRM.

Podsumowując, mimo przytoczonych wielu przykładów ataków na systemy TRM, a także bogactwa literatury dotyczącej tego zagadnienia, według najlepszej wiedzy autora, nie został opracowany spójny zbiór obejmujący wszystkie ataki na systemy TRM. Opisy zidentyfikowanych do tej pory ataków zostały dokonane na różnym poziomie szczegółowości i są rozproszone w wielu pracach. Warto także podkreślić, że jednoczesne zastosowanie wielu ataków i mierzenie osiągniętych w ten sposób efektów, jest także uważane za interesujący temat badań [69]. Akcentowana jest także możliwość tworzenia ataków adaptacyjnych, tzn. dostosowujących się do aktualnych warunków panujących w środowisku.

3.3.2. Taksonomie ataków

W literaturze dostępne są przykłady publikacji starających się dokonać klasyfikacji istniejących ataków przeciwko systemom TRM. W szczególności, artykuł [10] przytacza istotne cechy ataków, które można wyróżnić w celu stworzenia ich taksonomii (poniżej zaprezentowano najistotniejsze z tych cech z punktu widzenia tej rozprawy). Omawiane źródło zawiera także zestawienie możliwych do zastosowania mechanizmów zapobiegania poszczególnym atakom i prezentuje analizę, które z nich są skuteczne. Jednym z wniosków jest stwierdzenie, że zapobieganie jednemu typowi ataku może sprawić, że system stanie się

bardziej podatny na inne typy ataków, dlatego nie należy analizować systemu jedynie w kontekście odporności na określony, specyficzny rodzaj ataku. Praca wskazuje przykłady ataków dokonując ich klasyfikacji, nie stara się jednak zaprezentować wszystkich możliwych ataków i mechanizmów obronnych. Jedna z wcześniejszych prac autora rozprawy [79] także przywołuje, znalezione w literaturze, przykłady klasyfikacji ataków na systemy TRM, m.in. pod kątem poziomu wiedzy atakujących i celu ataku [10], mechanizmu działań adwersarzy [9], czy intencji atakujących i kosztów ataku [74].

W publikacji [57] zaprezentowano ogólną klasyfikację ataków, obejmującą takie aspekty, jak to, czy adwersarze uczestniczą w środowisku (atakujący wewnętrzny versus atakujący zewnętrzny), a także to, czy dany atak może być udaremniony przez wykorzystanie systemu TRM, czy może osłabić działanie systemu TRM, lub czy jest bezpośrednio wymierzony w system TRM.

Warto także podkreślić, że dostępne klasyfikacje ataków, np. [9], [46], [80]–[82], nie uwzględniają bardziej wyrafinowanych ataków, także tych które zostały szeroko opisane w literaturze.

3.4. PRACE DOTYCZĄCE SPOSOBÓW OCENY ODPORNOŚCI SYSTEMÓW NA ATAKI

Podrozdział zawiera krótką charakterystykę prac, które dotyczą zagadnienia stworzenia narzędzia, metodyki lub miar służących do oceny systemów zarządzania zaufaniem i reputacją.

3.4.1. Prace dotyczące metod oceny odporności systemów TRM na ataki

Analiza teoretyczna odporności systemów TRM na ataki jest stosowana m.in. w pracach [46], [80], [81]. Przyjmuje ona założenie, że przez weryfikację sposobu funkcjonowania danego systemu TRM będzie możliwe określenie, na jakie ataki jest on podatny. Analiza teoretyczna musi być stosowana w oparciu o szereg założeń upraszczających, które w praktyce mogą wpływać na wiarygodność wyników.

Niektóre artykuły starają się na bazie analizy teoretycznej dokonać porównania systemów zarządzania zaufaniem i reputacją, również pod kątem ich odporności na ataki [1], [9], [45]. Artykuł [83] autora rozprawy wskazuje, że: *„głównym problemem większości publikacji stosujących tę metodę jest brak precyzyjnego określenia, jakie kryteria są brane pod uwagę przy klasyfikowaniu systemu jako odporny lub podatny na dany rodzaj ataku, oraz określenia w jaki sposób można wykonać taką analizę dla innych systemów (lub ataków). Brak*

jasno zdefiniowanej metodyki czyni dokonaną analizę nieweryfikowalną i uniemożliwia określenie jej rzetelności. Wobec tego powtórzenie, rozszerzenie lub weryfikacja wyników zaprezentowanych przez autorów wiąże się z poważnymi trudnościami. Drugim ograniczeniem w porównywaniu dokonanych w ten sposób klasyfikacji systemów jest częste branie pod uwagę innego zbioru ataków lub definiowanie ataków w odmienny sposób. Nawet w przypadku wyeliminowania wyżej wymienionych problemów metodologicznych, stosowanie analizy teoretycznej wiąże się z:

- *ograniczeniem analizy jedynie do pewnego podzbioru ataków,*
- *brakiem oceny odporności systemów TRM w kategoriach ilościowych (ocena może zostać wykonana tylko w kategoriach jakościowych – tzn. czy atak będzie miał wpływ na funkcjonowanie środowiska, czy też nie).”*

Istotnym podejściem w kontekście oceny odporności systemów TRM na ataki jest ocena mechanizmów obrony przed atakami, które te systemy implementują. Za pracą [71] autora rozprawy warto zauważyć, że „*w literaturze są dostępne artykuły prezentujące propozycje mechanizmów wykrywania ataków bazujących na kooperacji przeciwko systemom zarządzania zaufaniem, jednakże ich zakres stosowania jest zwykle ograniczony. Dla sieci P2P został stworzony algorytm wykrywania potencjalnie złośliwych agentów współpracujących ze sobą* [84]*, ale posiada on istotne wady. Przede wszystkim może być stosowany jedynie w przypadku systemów zarządzania zaufaniem w sieciach P2P, a co więcej, wykrycie grupy współpracujących złośliwych agentów może nastąpić jedynie po fakcie wyświadczenia usługi o złej jakości (np. w przypadku sieci P2P – po pobraniu pliku o złej jakości). Co najistotniejsze, przedstawiony sposób wykrywania ataków kooperacyjnych działa jedynie w przypadku najprostszego ataku, w którym to złośliwe agenty zawyżają wartość zaufania do siebie nawzajem lub zaniżają wartość zaufania do rzetelnych agentów. Bardziej wysublimowane typy ataków nie zostaną wykryte. W przypadku ataku kooperacyjnego przedstawionego w* [45]*,* [85] *proponowana metoda okaże się przeciwskuteczna – rzetelne agenty zostaną zidentyfikowane jako złośliwe. Został opracowany także inny mechanizm wykrywający powiązania pomiędzy potencjalnie złośliwymi agentami* [86]*, jednak jego zastosowanie, podobnie jak w przypadku mechanizmu opisanego powyżej, ogranicza się praktycznie jedynie do sieci P2P.*[13]*”*

Jak zauważono w artykule [71]: *„publikacje próbujące w sposób kompleksowy oceniać odporność systemów zarządzania zaufaniem i reputacją na ataki, skupiają się zwykle na stworzeniu katalogu ataków i badaniu odporności na nie, a nie na próbie wykrywania wszelkich*

[13] gdyż bierze pod uwagę tylko agenty znajdujące się w bezpośrednim sąsiedztwie oceniającego agenta

podejrzanych zachowań, czy też niedozwolonej współpracy. Za przykład mogą świadczyć dwie prace prezentujące takie podejście: w artykule [87] *został przedstawiony katalog możliwych ataków (ale bez ich wyczerpujących opisów), w nieco szerszym ujęciu ataki tego typu zostały przedstawione w artykule* [88]. *Tego typu działania są pomocne ale niewystarczające, ponieważ istnieje znaczne prawdopodobieństwo, że zostanie stworzony nowy atak (np. bazujący na kooperacji złośliwych agentów), który może doprowadzić do kompromitacji systemu".*

Za pracą [83] autora rozprawy można stwierdzić, że: *„badania symulacyjne są najczęściej stosowaną metodą weryfikacji odporności systemów TRM na ataki. Poszczególne opracowania wykorzystujące tę metodę odmiennie definiują mierzone parametry, czy charakter wykonywanych symulacji. Co niemniej istotne, badania poszczególnych autorów biorą pod uwagę różne warunki panujące w środowisku (np. topologię sieci) i różny podzbiór ataków. Z tego względu badania wykonywane przez różnych autorów, w odniesieniu najczęściej do stworzonych przez siebie systemów, nie mogą być w łatwy sposób porównywane. W celu próby autorytatywnego wyboru obiektywnie lepszego systemu TRM do pewnych zastosowań, konieczne jest przeprowadzenie własnych kompleksowych badań. Pewnym przełomem, wartym podkreślenia, jest stworzenie dedykowanych narzędzi symulacyjnych ułatwiających wzajemne porównywanie różnych systemów TRM."* Przykłady takich narzędzi zawarto w podrozdziale 3.4.3. Wykorzystanie aktualnie dostępnych narzędzi wiąże się jednak z istotnymi ograniczeniami.

Zwykle metody symulacyjne używane są do analizy odporności systemów TRM jedynie w kontekście opisanych szeroko w literaturze i zwykle najprostszych ataków. Jak zauważa autor rozprawy w artykule [83]: *„na tym tle wyróżnia się publikacja* [89]*, w której zostało użyte pojęcie „worst case attack" sugerujące rozważanie najbardziej efektywnego ataku dla danego systemu zarządzania zaufaniem i reputacją. Mimo to, nie został opisany sposób wyboru (lub konstrukcji) ataku, który doprowadziłby do największego spadku efektywności systemu. W pracy tej znajdują się także wyniki badań symulacyjnych w odniesieniu do kilku typów ataków. Z tego względu można założyć, że wybór ataku, który jest najbardziej skuteczny, wykonywany jest a posteriori na bazie symulacji znanych ataków, co właściwie niczym nie różni się od zwykłych badań symulacyjnych. Niemniej omawiane źródło zasługuje na specjalne potraktowanie ze względu na użycie pojęcia najbardziej efektywnego ataku dla danego systemu TRM."*

Cechą wspólną dla większości badań systemów TRM wykonanych w literaturze jest stosowanie własnej metodyki ich przeprowadzania, w szczególności poprzez zdefiniowanie, nie zawsze w sposób precyzyjny, własnych miar efektywności systemów [69], [90], [91]. Opis

wybranych propozycji takich miar zawarto w podrozdziale 3.4.2. W wielu przypadkach charakter badań jest niewystarczający do kompleksowej oceny wiarygodności badanego systemu TRM i niejednokrotnie służy podkreśleniu korzystnych właściwości danego systemu TRM.

Szczególnie istotną kwestią jest brak kompleksowej metodyki oceny wiarygodności systemów TRM, np. autor monografii [8] podkreśla, że jakkolwiek wiele systemów zarządzania zaufaniem było proponowanych w literaturze, to wciąż nie ma szeroko akceptowalnej metody ewaluacji takich systemów. Jednocześnie przedstawia on pewne idee z tym związane. Proponuje, aby w celu ewaluacji systemów zarządzania zaufaniem stworzyć w sposób analityczny specjalne scenariusze, w celu oceny jak bardzo wyniki uzyskiwane za pomocą poszczególnych systemów zarządzania zaufaniem odbiegają od prawidłowych wartości zaufania i ryzyka (przy czym nie definiuje w jaki sposób tę prawdziwą wartość wyznaczyć). Zdaniem autora jednym z najistotniejszych aspektów zarządzania zaufaniem i reputacją jest ocena odporności na działania adwersarzy. Postuluje, że ze względu na to, że wiele metod zarządzania zaufaniem stosuje regułę większości, to są one podatne na działania wielu współpracujących adwersarzy lub adaptacyjnych ataków, takich jak dyskryminacja. Wskazuje też modele adwersarzy, które dotyczą aspektów takich jak: wiedza adwersarza, cele adwersarzy, stopień współpracy adwersarzy, zasoby adwersarza, czy złożoność adwersarza. W niniejszej rozprawie poniekąd rozwinięto ideę zaproponowaną przez autora monografii [8], przy czym nie wszystkie wymienione aspekty modelu adwersarza uznano za jednakowo istotne w kontekście ataków złośliwych agentów, dodano także do nich inne (modele ataków zostały omówiony w podrozdziale 4.3).

Innymi istotnymi aspektami poruszanymi w publikacji [8] są wymagania dotyczące uwierzytelnienia. Ma to szczególne znaczenie dla wszystkich strategii ataków wykorzystujących ataki fałszowania tożsamości, w szczególności ataku kreacji wielu tożsamości (ang. „Sybil attack”).

Publikacja [92] jest interesującą pozycją, ale dotyczy nieco innego problemu – uczciwego podziału zasobów. Mimo to część z zawartych wniosków może być wykorzystana przy opracowaniu metodyki oceny wiarygodności systemów TRM. W pracy dokonano przeglądu wybranych koncepcji związanych z reputacją i zaufaniem w systemach teleinformatycznych. Szczególny nacisk położono na maksymalizację użyteczności działań agentów. Jest to poniekąd zbieżne z zagadnieniami dotyczącymi uczciwości agentów, poruszanymi niejednokrotnie w literaturze. W dokumencie przedstawiono pojęcia mechanizmu budowy reputacji i zaufania (ang. „Reputation and Trust Building Scheme”) i silnika agregacji

danych reputacyjnych (ang. „Reputation data Aggregation Engine"). Podjęte zostały także rozważania na temat sposobu porównania efektywności systemów wieloagentowych, które są pozbawione mechanizmów zarządzania zaufaniem i reputacją, z systemami, które implementują takie mechanizmy. Słuszną uwagą autora jest stwierdzenie, że w przypadku wielu badań porównanie odbywa się na podstawie globalnej wydajności systemu przy założeniu, że w obydwu przypadkach agenty postępują w ten sam sposób (mają te same polityki), co jest założeniem naiwnym.

Za pracą [92] można rozważać także motywacyjną zgodność agentów do działań rzetelnych, w celu oceny systemów TRM. W uproszczeniu można stwierdzić, że działanie systemu TRM jest skuteczne wtedy gdy zapewnia motywacyjną zgodność działań agentów z celami działania środowiska jako całości. W tym kontekście badania [92] wprowadzają istotny przełom, gdyż wskazują wymagania stawiane przed systemami TRM w celu zapewnienia ochrony przed działaniami samolubnymi. W praktyce jednak wydaje się, że takie podejście nie zawsze jest wystarczające do oceny odporności systemów na ataki, gdyż nie bierze ono pod uwagę agentów złośliwych, zdeterminowanych do obniżenia efektywności działania systemów TRM, nawet w przypadku konieczności poniesienia pewnych kosztów.

3.4.2. Propozycje miar ocen efektywności systemów TRM

W tym punkcie zaprezentowano krótki przegląd najważniejszych miar efektywności systemów zarządzania zaufaniem i reputacją, które zostały określone w literaturze.

W artykule [69] zdefiniowano poziom wykrywania złośliwych agentów (ang. „malicious node detection performance" – MDP), który reprezentuje średni poziom detekcji złośliwych agentów (tzn. ile średnio zostało odkrytych złośliwych agentów w środowisku):

$$MDP = \frac{\sum_{i \in M} n_B^i}{|M|}$$

gdzie: n_B^i - liczba rzetelnych agentów, które wykryły, że agent i jest złośliwy; M – zbiór złośliwych agentów; B – zbiór rzetelnych agentów.

Analogiczną miarą, także zdefiniowaną w omawianym źródle, jest poziom fałszywych alarmów:

$$FAR = \frac{\sum_{i \in B} n_B^i}{|B|}$$

Podobna miara efektywności (trafność detekcji) została także zdefiniowana w pracy [35].

W artykule [69] zdefiniowano poziom dostarczonych pakietów (ang. „packet delivery ratio") jako stosunek pakietów dostarczonych do wysłanych (w pracy tej jest opisane użycie systemu TRM dla protokołów routingu). Warto podkreślić, że także w innych pracach (np. w [45]) stosowane są podobne miary, określane często jako efektywność systemu lub efektywność sieci (ang. „network effectiveness"), która może wbrew nazwie być stosowana także w innych typach środowisk. W artykule [91] zdefiniowano analogiczną miarę – poziom strat pakietów (ang. „packet loss") jako procent pakietów niedostarczonych spośród pakietów wysłanych.

W publikacjach [91], [35] mierzono zużycie energii przez agenty, będące węzłami sieci WSN, podczas gdy był uruchomiony system zarządzania zaufaniem i porównywano je z poziomem zużycia energii w sytuacji braku użycia systemu zarządzania zaufaniem i reputacją – na tej podstawie wyznaczano miarę zużycia energii (ang. „energy consumption") z uwagi na użycie systemu zarządzania zaufaniem.

Powyższe przykłady miar efektywności systemów zarządzania zaufaniem mogą być zastosowane w przypadku wybranych środowisk. Odmiennym podejściem jest próba stworzenia miar, które mogłyby znaleźć zastosowanie do szerokiej klasy systemów zarządzania zaufaniem i reputacją. Przykładem pracy proponującej ogólne miary efektywności systemów TRM jest artykuł [93] autora niniejszej rozprawy. W artykule tym zaproponowano między innymi miary takie jak: efektywność systemu, odporność systemu, zysk efektywności systemu, zysk absolutny efektywności systemu, szybkość propagacji informacji o zmianie zachowania, narzut obliczeniowy na decyzję, narzut obliczeniowy na obsługę żądań, narzut na zasoby pamięci, narzut na energię, reputacja całkowita agentów złośliwych, reputacja całkowita agentów rzetelnych i inne. Część z zaproponowanych miar dotyczy szerszego zagadnienia oceny funkcjonowania systemów TRM, a nie tylko oceny ich odporności na ataki. Miary, które są istotne w kontekście oceny wiarygodności systemów TRM, zostały rozszerzone i szerzej opisane w rozdziale 5.

3.4.3. Opis symulatorów służących do oceny wiarygodności systemów TRM

Przegląd prac dotyczących narzędzi został dokonany przez autora rozprawy w artykule [93]: *„TRMSim-WSN* [67] *jest symulatorem działania systemów zarządzania zaufaniem i reputacją, stworzonym w języku Java* [94] *i udostępnionym na zasadach wolnego oprogramowania. Zadaniem, jakie przyświecało autorom tego narzędzia, było umożliwienie łatwej implementacji kolejnych systemów zarządzania zaufaniem. Jest to narzędzie*

ukierunkowane do wykonywania symulacji działania systemów zarządzania zaufaniem, ale jego istotnym ograniczeniem ze względu na przyjęte założenia jest to, że nadaje się do symulowania działania takich systemów tylko w specyficznych typach środowisk, przede wszystkim w bezprzewodowych sieciach sensorowych.

TRMsim-WSN jest najbardziej rozbudowanym narzędziem symulacyjnym spośród opisanych w literaturze i analizowanych przez autora niniejszej pracy. Posiada jednak trzy istotne z perspektywy celu niniejszej rozprawy ograniczenia. Po pierwsze, jest dedykowany dla sieci WSN, w związku z czym nie nadaje się do przeprowadzania symulacji dla innych środowisk zastosowań systemów zarządzania zaufaniem i reputacją, np. e-handlu. Narzędzie to zakłada, że agenty mogą posiadać informacje pochodzące z pośrednich obserwacji interakcji dokonywanych przez inne agenty (ze względu na bezprzewodowy charakter komunikacji w sieciach WSN), co ogranicza zakres zastosowań systemów zarządzania zaufaniem i reputacją. Po drugie, nie zawiera implementacji znacznej liczby znanych ataków przeciwko systemom zarządzania zaufaniem i reputacją. Zaimplementowany jest tylko atak oscylacji zachowania (on-off) oraz atak stały (constant), natomiast implementacja bardziej wyrafinowanych ataków, ze względu na konstrukcję narzędzia, nie jest łatwa, aczkolwiek możliwa. Po trzecie, w tym symulatorze nie są analizowane wszystkie istotne parametry związane z działaniem systemów zarządzania zaufaniem w odniesieniu do odporności na ataki, a prezentowane statystyki ograniczają się w praktyce do monitorowania stosunku interakcji zakończonych sukcesem do wszystkich nawiązanych interakcji, co wydaje się niewystarczające z punktu widzenia analizy wszystkich możliwych ataków.

Interesującym narzędziem jest także symulator ART [95], [96]. *Narzędzie to w założeniu ma służyć jako platforma, na bazie której różne zespoły badawcze mogą porównywać stworzone przez siebie systemy zarządzania zaufaniem i reputacją z użyciem predefiniowanych metryk oraz jako narzędzie pozwalające przeprowadzać powtarzalne eksperymenty w celu oceny poprawności przyjętych założeń.*

Autorzy [95], [96] w trakcie tworzenia narzędzia zdefiniowali następujący problem badawczy, który może zostać rozwiązany z użyciem dedykowanych do tego celu systemów zarządzania zaufaniem i reputacją, których skuteczność może zostać przetestowana w narzędziu. Rozważany jest rynek obrazów, na którym występuje wielu rzeczoznawców posiadających zróżnicowany poziom kompetencji w różnych malarskich stylach i erach. Klienci rzeczoznawców zamawiają wyceny dla obrazów. Jeżeli dany ekspert nie posiada wystarczających kompetencji, aby wykonać wycenę, może zakupić taką usługę od innego rzeczoznawcy. Rzeczoznawcy oceniają trafność własnych wycen poprzez cenę jaką oferują za

jej wykonanie. Rzeczoznawcy mogą kłamać (np. starać się zawyżyć własny poziom kompetencji). Rzeczoznawcy generują wyceny z użyciem własnej wiedzy oraz wycen dostarczonych przez innych rzeczoznawców. Im trafniejsze są wyceny rzeczoznawców dostarczane dla klientów, tym mają oni więcej klientów, co przekłada się na zyski. Rzeczoznawcy mają możliwość także zakupienia od innych rzeczoznawców informacji o ocenach kompetencji innych rzeczoznawców (czyli zakupienia rekomendacji). W trybie gry (używanym do porównywania własnych systemów zarządzania zaufaniem pomiędzy różnymi zespołami badawczymi) zwycięża ten rzeczoznawca, który posiada największe saldo na koncie.

Narzędzie to może być przydatne w przypadku próby tworzenia systemów zarządzania zaufaniem dla specyficznego celu (jako swoisty konkurs pomiędzy różnymi zespołami badawczymi), ale nie jest użyteczne w przypadku próby oceny systemów zarządzania zaufaniem i reputacją stworzonych dla innych zastosowań. Wydaje się, że problem postawiony przez autorów tego narzędzia jest zbyt abstrakcyjny i przez to narzędzie staje się nieprzydatne w praktycznych zastosowaniach. Ograniczenia w użyciu narzędzia ART zostały dostrzeżone przez wielu autorów np. [97], [98].

Kolejnym przykładem narzędzia symulacyjnego jest TREET [99]. *Służy ono do oceny systemów zaufania i reputacji używanych w e-handlu. Mimo tego, że nie można wykorzystać tego narzędzia do ewaluacji szerokiej klasy systemów zarządzania zaufaniem i reputacją, to należy podkreślić, że stopień szczegółowości i kompleksowości osiągnięty przez autorów zasługuje na uwagę. Dla systemów zarządzania zaufaniem w e-handlu jest to najbardziej dojrzały z dostępnych symulatorów, ale jego zastosowanie jest niestety ograniczone tylko do tej klasy systemów i nie może być, w łatwy sposób, rozbudowane (ze względu na zdefiniowane scenariusze testów, specyficzne dla systemów e-handlu). Dodatkowo, mimo że kod narzędzia został udostępniony, to ograniczenia we wprowadzaniu zmian i tworzeniu własnych komponentów sprawiają, że jego użycie w znacznej liczbie przypadków nie będzie możliwe.*

Podobnym narzędziem symulacyjnym, także służącym do ewaluacji systemów zarządzania zaufaniem i reputacją wykorzystywanych w e-handlu, jest narzędzie przedstawione w pracy [100]. *Na uwagę zasługuje także symulator EStarMom* [101], *posiadający jednak podobne ograniczenia jak powyżej przedstawione symulatory."*

Inną publikacją przedstawiającą narzędzie do oceny odporności systemów TRM na ataki w środowisku e-handlu jest [102]. Zdaniem autorów istniejące narzędzia opierają się głównie na symulacjach, podczas gdy to stworzone przez nich może wykorzystywać zarówno symulacje jak i dane pozyskane z rzeczywistych środowisk. Autorzy wskazują również inne słabości istniejących badań dotyczących systemów TRM. Podkreślają między innymi, że

propozycje nowych systemów TRM w większości są oceniane na bazie własnych symulatorów ich autorów i porównywane z bardzo niewieloma innymi systemami, a co więcej, większość narzędzi opiera się na pojedynczym scenariuszu symulacyjnym, co nie może być postrzegane jako rzetelne. Narzędzia, takie jak ART lub TREET są głównie oparte na symulacjach i nie mogą odzwierciedlić rzeczywistych właściwości konkretnego środowiska. Poza tym nie zostały one stworzone z myślą o ewaluacji odporności systemów zarządzania zaufaniem i reputacją na ataki polegające na nierzetelnych ocenach. Także często te środowiska używają jedynie pojedynczej metryki w celu oceny modeli zaufania. Zdaniem autorów publikacji [102], unikalność proponowanego narzędzia polega na tym, że zapewnia różnorodność przeprowadzanych symulacji i badań oraz umożliwia swobodne określanie parametrów eksperymentów. Dodatkowo narzędzie potrafi oceniać odporność systemów TRM na nierzetelne opinie agentów oraz wspiera wiele systemów zarządzania zaufaniem i reputacją oraz typów ataków (zarówno prostych, jak również bardziej złożonych). Zaimplementowano w narzędziu następujące systemy TRM: BRS, TRAVOS, iCLUB, BLADE, WMA, ProbCog, Personalized, eBay oraz następujące ataki[14]: Constant; Camouflage; Whitewashing; Sybil; Sybil Camouflage; Sybil Whitewashing. Autorzy wskazują także na użycie kilku miar odporności systemów TRM na ataki, ale nie przedstawiają ich definicji. Przyszłe prace autorów mają polegać na rozszerzeniu narzędzia w celu uwzględnienia systemów używanych w innych środowiskach niż e-handel. Główną wadą artykułu jest bardzo pobieżny opis stworzonego środowiska, oraz to że kody źródłowe narzędzia nie są dostępne, wobec czego trudno ocenić rzeczywistą przydatność narzędzia do oceny wiarygodności systemów TRM.

Monografia [8] także wskazuje symulator do oceny efektywności systemów TRM funkcjonujących w ramach platform e-handlu. W ramach symulacji każdy agent był charakteryzowany przez następujące parametry:

- prawdopodobieństwo, że agent wyśle raport (ocenę) interakcji, jeżeli jej wynik był pozytywny;
- prawdopodobieństwo, że agent wyśle raport, jeżeli wynik interakcji był negatywny;
- wybrana strategia zachowania;
- próg reputacji używany przez niektóre strategie;
- prawdopodobieństwo oszustwa, będące parametrem niektórych strategii.

Zwykle w ramach symulacji występowały dwie grupy agentów: rzetelne i nierzetelne, w obydwu grupach agenty miały taką samą charakterystykę. Mierzona była reputacja

[14] Nie zostało to określone jednoznacznie, jednakże te ataki są w pracy wspominane.

poszczególnych agentów. Istotne jest to, że w tych symulacjach strategie agentów koncentrowały się jedynie na wykonywaniu akcji, a nie na fałszowaniu rekomendacji, co jest wszakże często głównym składnikiem ataku. Z tego względu także ten symulator nie wydaje się być wystarczająco ogólny do przeprowadzania badań różnorodnych systemów TRM.

Autor rozprawy, w swojej wcześniejszej publikacji [93] także zaznaczał, że: *„wielu autorów prac, opisujących nowe systemy zarządzania zaufaniem i reputacją, dokonuje ewaluacji stworzonych przez siebie propozycji za pomocą własnych narzędzi symulacyjnych. Jest to spowodowane dwoma faktami: po pierwsze, nie istnieje narzędzie zaakceptowane przez szerokie środowisko badawcze, a po drugie, wiele systemów zarządzania zaufaniem różni się na tyle znacznie (korzystają z różnych danych wejściowych, generują różne dane wyjściowe i reagują na różne zdarzenia), że stworzenie jednolitego środowiska symulacyjnego jest trudne* [100]. *Przykładami takich narzędzi mogą być: symulator przeznaczony dla sieci peer-to-peer* [103], *czy symulator stworzony dla potrzeb sieci WSN* [11]. *Cechą wspólną tego typu symulatorów jest to, że są przeznaczone tylko dla określonego typu środowiska, w którym są wykorzystywane systemy zarządzania zaufaniem i reputacją oraz to, że najczęściej biorą pod uwagę tylko najprostszy sposób zachowania się złośliwych agentów – czyli stosowanie ataku stałego* [1]. *Warto także podkreślić, że niejednokrotnie do oceny efektywności tworzonych systemów TRM są wykorzystywane rozbudowane środowiska symulacyjne, takie jak OPNET* [104], *czy NS2* [105]. *Wadą takich rozwiązań jest to, że w takich przypadkach nie są wykorzystywane specyficzne metryki dla systemów zarządzania zaufaniem i reputacją oraz to, że symulacje przeprowadzane są tylko w określonym specyficznym typie środowiska (np. sieci WSN, czy MANET), w związku z czym nie są one możliwe do wykorzystania do porównywania różnych systemów zarządzania zaufaniem i reputacją. Dodatkowo użycie tego typu symulatorów może wiązać się ze znacznym narzutem ze względu na użycie pełnego stosu protokołów, co nie zawsze jest konieczne w przypadku oceny odporności i efektywności systemów zarządzania zaufaniem."*

3.5. PODSUMOWANIE PRZEGLĄDU STANU WIEDZY

Mimo, że tematyka systemów zarządzania zaufaniem i reputacją, a także ataków na takie systemy, jest szeroko eksplorowana, to wśród zespołów badawczych wydaje się przeważać pogląd, że wciąż brak jest określenia sposobu wyczerpującej oceny wiarygodności tego typu systemów. W szczególności, jak autor rozprawy zaznaczał we wcześniejszej

publikacji [106], opisane w literaturze badania odporności systemów TRM na ataki charakteryzuje:

- *„brak precyzyjnie zdefiniowanej metodyki przeprowadzania badań, w celu uzyskania możliwości porównywania wyników dla różnych systemów;*
- *ograniczenie tylko do wybranego zestawu miar, np. efektywności, bez uwzględnienia tego jak zdobycie wysokiego zaufania lub reputacji przez złośliwe agenty może przyczynić się do zaburzenia efektywności w dłuższej perspektywie;*
- *nie zawsze kompleksowy charakter przeprowadzanych badań, które potwierdzają skuteczność proponowanego systemu, a ukrywają jego wady czy ograniczenia;*
- *brak wykorzystania ogólnodostępnych narzędzi symulacyjnych, które umożliwiałyby zarówno łatwe dodawanie nowych systemów TRM, jak i ataków."*

Najczęściej badanymi atakami, co nie jest w żaden sposób zaskakujące, są te najprostsze, tzn. atak oscylacji zachowania czy atak stały. Dość często badane są ataki polegające na fałszowaniu rekomendacji (wychwalanie i oczernianie). W pojedynczych przypadkach są badane bardziej wysublimowane ataki, w szczególności te oparte na kooperacji agentów. Zwykle do eksperymentalnego badania ataków (najczęściej w postaci symulacji) stosowana jest ogólna miara określająca efektywność systemu. Definicje tej miary różnią się w zależności od autorów badań, jednak zwykle jest to pewna suma wyników interakcji – im jest ona wyższa tym system działa efektywniej. Jest to uzasadnione faktem, że wysoka wartość tej miary sugeruje, że złośliwe agenty nie są wybierane jako partnerzy interakcji. Miara ta jest niewystarczająca, zdaniem autora niniejszej rozprawy, z kilku powodów:

- nie uwzględnia możliwości celowego zmanipulowania określonej grupy agentów,
- nie uwzględnia faktu, że celem złośliwych agentów może nie być dążenie do minimalizacji średniej efektywności systemu, ale mogą im przyświecać inne cele, np. możliwość nagłego spadku efektywności w określonym momencie czasowym,
- nie bierze pod uwagę na ile są zmanipulowane wartości zaufania lub reputacji pomiędzy poszczególnymi agentami. Zagadnienie to może być istotniejsze, po pierwsze z uwagi na to, że niekoniecznie celem agentów atakujących jest spadek efektywności, a poza tym, że osiągnięcie wyższych wartości zaufania do atakujących agentów może umożliwić im w przyszłości przeprowadzenie ataku.

Niektóre z powyższych wniosków zostały także wyciągnięte przez autorów niektórych teoretycznych opisów ataków, w szczególności w publikacji [73], ale zwykle nie są proponowane miary, które byłyby pozbawione wymienionych wad. Z tego względu,

w rozdziale 5 przedstawiono propozycje nowych miar, które będą pozwalały na dostrzeżenie innych celów atakujących i szerszą ocenę systemów TRM pod kątem działań adwersarzy.

Jak wykazano w tym rozdziale, zgodnie z najlepszą wiedzą autora, brak jest pracy podsumowującej różne propozycje w zakresie badań odporności systemów TRM na ataki, ale przede wszystkim brak jest szeroko akceptowanej metodyki oceny wiarygodności systemów TRM. Z tego względu w rozdziale 5 zostanie przedstawiona propozycja metodyki oceny wiarygodności systemów TRM.

4. MODEL ŚRODOWISKA, SYSTEMU TRM I ATAKU

Istnieje wiele systemów zarządzania zaufaniem i reputacją, wśród których, jak zauważono wcześniej, można wyróżnić wiele cech wspólnych. Celem rozdziału jest prezentacja modelu wystarczająco ogólnego, aby obejmował jak najwięcej systemów TRM, ale jednocześnie na tyle szczegółowego, aby na jego podstawie umożliwić przeprowadzenie nietrywialnych badań wiarygodności takich systemów, także w ujęciu ilościowym. Model systemu TRM został utworzony zarówno na bazie literatury jak i własnych wniosków autora i zaprezentowany w podrozdziale 4.2, w oparciu o model środowiska przedstawiony w podrozdziale 4.1. Modele ataku na systemy TRM zostały zaprezentowane w podrozdziale 4.3.

4.1. MODEL ŚRODOWISKA

Na mocy definicji 2, środowisko składa się z agentów wchodzących w interakcje polegające na świadczeniu określonych usług. Warto jednak bardziej szczegółowo określić cechy środowiska, poprzez charakterystykę jego poszczególnych elementów. Przedstawiony model środowiska jest prawidłowy zarówno w przypadku środowisk, w których działają systemy TRM, jak i środowisk, w których nie funkcjonuje taki system. Na cele dalszej pracy, przyjęto następujące założenia, określające model środowiska:

- w środowisku działają agenty ze skończonego zbioru A, przy czym $|A| = n, A = \{a_1, \dots, a_n\}$, gdzie: $a_1, \dots, a_n$ – agenty o numerach (identyfikatorach) odpowiednio $1, .., n$;
- podzbiorem zbioru agentów A jest zbiór agentów rzetelnych A_B, $(A_B \subseteq A)$, $|A_B| = n_B$;
- podzbiorem zbioru agentów A jest zbiór agentów złośliwych[15] A_M, $(A_M \subseteq A)$, $|A_M| = n_M$,
    - spełnione są także zależności: $A_B \cap A_M = \emptyset$, $A_B \cup A_M = A$, $n_B + n_M = n$;
- w środowisku mogą być świadczone usługi z określonego skończonego zbioru usług $U = \{u_1, \dots, u_l\}$;
- każdy z agentów $a_1, \dots, a_n$ może świadczyć usługi ze zdefiniowanego zbioru usług tego agenta, tj. $U_{a_1}, \dots, U_{a_n}$, przy czym każdy z tych zbiorów jest dowolnym podzbiorem zbioru U, tj. $\forall k \in [1, n]$: $U_{a_k} \subseteq U$;

[15] Obejmuje także agenty samolubne.

    - w szczególności, może istnieć pewien zbiór $A_X \subset A$ agentów, którzy nie świadczą żadnych usług, tzn. $\forall a_x \in A_X: U_{a_x} = \emptyset$,
    - w szczególności może istnieć pewien zbiór $A_Y \subset A$ agentów, którzy mogą świadczyć wszystkie usługi dostępne w środowisku, tzn. $\forall a_y \in A_Y: U_{a_y} = U$;
- suma zbiorów usług świadczonych przez każdego z agentów w środowisku daje zbiór usług świadczonych w środowisku, tj. $\bigcup_{i=1}^{n} U_{a_i} = U$;
- usługi w środowisku mogą być świadczone z jakością q o wartości ze zbioru wartości jakości usług Q;
    - zbiór Q może być nieskończony;
    - zbiór Q może zawierać takie wartości jakości q, które nie są aktualnie świadczone w środowisku;
    - wartości w zbiorze Q da się uszeregować od najniższej do najwyższej;
    - w zbiorze Q istnieje wartość jakości usługi odpowiadająca brakowi wyświadczenia usługi ($q_0 \in Q$);
    - w zbiorze Q istnieje wartość jakości usługi odpowiadająca jakości idealnej ($q_{max} \in Q$), taka że $\forall q_k \in Q, q_k \neq q_{max}: q_k < q_{max}$;
    - najniższa wartość jakości w zbiorze usług to wartość q_{min} taka że: $\forall q_k \in Q$, $q_k \neq q_{min}: q_k > q_{min}$;
    - w zbiorze Q mogą istnieć takie wartości q_k że $q_k < q_0$, tzn. że może istnieć taka jakość usługi, że jest ona gorsza niż brak usługi;
- każdy z agentów, dla każdego ze świadczonych przez niego usług ma określoną maksymalną jakość tej usługi, która może być dostarczona, np. maksymalna jakość usługi u_l, świadczonej przez agenta a_k wynosi $q_{a_k}^{u_l} = q$, gdzie q jest pewną wartością ze zbioru Q ($q \in Q$),
    - przyjmuje się, że agent który może świadczyć daną usługę z maksymalną jakością q jest w stanie także świadczyć tę usługę z jakością niższą od q, będącą w zbiorze Q;
- istnieje zbiór agentów – usługodawców A_P, który zawiera agentów, którzy świadczą przynajmniej jedną usługę, tzn. $\forall a_k \in A_P: U_{a_k} \neq \emptyset$, $A_P \subseteq A$, $A_P \neq \emptyset$
- dla każdej usługi u_l istnieje zbiór agentów – usługodawców tej usługi: $A_{P:u_l}$, taki że: $\forall a_k \in A_{P:u_l}: u_l \in U_{a_k}$;

- każdy z agentów $a_1, \dots, a_n$ może żądać usług ze zdefiniowanego zbioru usług żądanych przez tego agenta, tj. $U_{R:a_1}, \dots, U_{R:a_n}$, przy czym każdy z tych zbiorów jest dowolnym podzbiorem zbioru U, tj. $\forall k \in [1, n]$: $U_{R:a_k} \subseteq U$;
    - w szczególności, może istnieć pewien zbiór $A_{R:X} \subset A$ agentów, którzy nie żądają żadnych usług, tzn. $\forall a_x \in A_{R:X}$: $U_{R:a_x} = \emptyset$,
    - w szczególności może istnieć pewien zbiór $A_{R:Y} \subset A$ agentów, którzy mogą żądać wszystkich usług dostępnych w środowisku, tzn. $\forall a_y \in A_{R:Y}$: $U_{R:a_y} = U$;
- istnieje zbiór agentów – usługobiorców A_R, który zawiera agentów, którzy przynajmniej raz zażądają świadczenia co najmniej jednej usługi, tzn. $\forall a_k \in A_R$: $U_{R:a_k} \neq \emptyset$, $A_R \subseteq A$
- suma zbiorów usług żądanych przez każdego z agentów w środowisku daje podzbiór zbioru usług świadczonych w środowisku, tj. $\bigcup_{i=1}^{n} U_{R:a_i} \subseteq U$,
    - czyli nie wszystkie usługi świadczone w danym środowisku muszą być żądane[16];
- w ogólnym przypadku wszystkie powyższe parametry są zmienne, tzn. mogą się zmieniać w czasie, każdy z powyższych parametrów może być wyznaczony w określonym momencie działania środowiska – czyli w czasie m,
    - w dalszej części rozprawy przyjęto, że parametry te nie zmieniają się w czasie i są ustalone dla danego środowiska[17], chyba że jawnie wskazano inaczej.

4.1.1. Topologia środowiska

Istotną charakterystyką środowiska jest jego topologia rozumiana jako struktura połączeń (relacji) pomiędzy agentami. Połączenie pomiędzy agentami umożliwia agentom nawiązanie interakcji (żądanie i świadczenie usługi). Dla danego środowiska można wyróżnić trzy warstwy topologii połączeń:

- połączenia komunikacyjne – odpowiadają połączeniom w warstwie przekazywania komunikatów (np. pakietów sieciowych);
- połączenia w warstwie usługowej – agenty, które posiadają połączenie w warstwie usługowej mogą wymieniać się usługami;

[16] W praktyce, wszystkie usługi mogą być żądane, tzn. zwykle będzie zachodzić $\bigcup_{i=1}^{n} U_{R:a_i} = U$, ale nie jest to konieczne

[17] Ograniczenie to nie wynika z kłopotów odzwierciedlenia tego faktu w modelu, ale głównie z problemów związanych z oznaczeniami, bowiem dla każdego parametru konieczne byłoby dodanie wymiaru czasu – m, co wydaje się, że niepotrzebnie skomplikowałoby dalsze rozważenia, nie wprowadzając w zamian istotnego ulepszenia modelu.

- połączenia w warstwie rekomendacyjnej – połączenia te istnieją tylko jeżeli w środowisku działa system zarządzania zaufaniem i reputacją – agenty, które posiadają połączenie w warstwie rekomendacyjnej mogą przekazywać sobie rekomendacje lub oceniać wzajemnie zaufanie lub reputację.

W ogólnym przypadku topologia środowiska pod kątem każdego typu połączeń może być odmienna. W szczególności połączenie w warstwie usługowej nie musi odpowiadać komunikacyjnemu połączeniu pomiędzy agentami (w skrajnym przypadku zarówno żądanie usługi, jak i sama usługa może być świadczona w oparciu o niezależne sieci). Dobrym przykładem jest środowisko będące platformą e-handlu, gdzie agentami są klienci i sprzedający, którzy wymieniają się produktami (usługami), przy czym każdy agent może żądać usługi od każdego agenta oferującego ją, a także dokonywać oceny zaufania do każdego innego agenta. Żądanie usługi jest przekazywane na bazie niezależnej sieci (samej platformy e-handlu w oparciu np. o Internet), natomiast dostarczenie produktu jest wykonywane w oparciu o operatorów pocztowych (co może być utożsamiane z kolejną niezależną siecią połączeń komunikacyjnych). W przypadku niektórych typów środowisk bez żadnej straty można pominąć warstwę połączeń komunikacyjnych, gdyż nie będzie ona miała istotnego znaczenia w analizie efektywności działania systemów zarządzania zaufaniem i reputacją[18].

Interesującym przykładem środowiska, obok wskazanej wcześniej platformy e-handlu, gdzie także występuje połączenie pomiędzy dowolną parą agentów w warstwie usługowej, jest sieć P2P. Przykładami takich środowisk są także sieci WSN i MANET, w których topologia sieci jest znacznie bardziej skomplikowana, a połączenia w warstwie usługowej są odzwierciedleniem połączeń w warstwie komunikacyjnej. Dodatkowo w sieciach MANET ta topologia może się zmieniać w czasie.

4.1.2. Żądanie świadczenia usługi

W ramach działania środowiska, przyjmujemy, że w kolejnych momentach czasowych $m_1, m_2, m_3, \ldots$ pojawiają się żądania usługobiorców. Kolejne żądania można uszeregować pod kątem czasu pojawienia się oraz po identyfikatorze agenta i ponumerować spójnie w całym

[18] Przypadkiem, w którym warstwa komunikacyjna połączeń może mieć znaczenie jest sytuacja, w której agenty monitorują (obserwują) komunikaty sieciowe wysyłane przez inne agenty i uwzględniają te obserwacje przy ocenie zaufania lub reputacji innych agentów. W szczególności, jeżeli usługą jest przekazywanie komunikatu sieciowego, to rozważenie warstwy komunikacyjnej ma duże znaczenie, ale wynika to z faktu, że w takim przypadku połączenia w warstwie usługowej są powieleniem połączeń w warstwie komunikacyjnej.

systemie, dzięki temu możliwe jest osiągnięcie spójności sekwencyjnej[19]. Przyjmujemy, że interakcja[20] o numerze k (czyli interakcja i_k) rozpoczęła się w momencie czasowym m_k. Dodatkowo przyjmujemy, że analiza działania środowiska rozpoczyna się w czasie $m_0 = 0$, a pierwsza interakcja zostaje zapoczątkowana w czasie $m_1 > 0$.

Warto podkreślić, że rozważane są tylko te momenty czasowe, w których pojawiają się żądania usługobiorców (a dokładniej momenty, w których usługobiorcy zidentyfikują potrzebę skorzystania z pewnej usługi)[21]. Co oczywiste, pojawianie się kolejnych żądań nie musi odbywać się w stałych interwałach czasowych, tj. $\forall k\colon\ m_k - m_{k-1} \neq const$.

Przyjmujemy także, że każda z interakcji ma pomijalny czas trwania. W związku z tym nie może zdarzyć się sytuacja, że kolejna interakcja rozpocznie się, zanim poprzednia zostanie zakończona. W danym momencie trwa co najwyżej jedna interakcja.[22] Warto podkreślić że jest to istotne ograniczenie modelu, szczególnie w przypadku niektórych typów środowisk. Szersza dyskusja tego zagadnienia znajduje się w punkcie 4.1.7.

Przyjmijmy, że rozważamy działanie środowiska w czasie od m_0 do m_{end} i w tym czasie zostało zapoczątkowanych l_I interakcji ($l_I \in \mathbb{N}$). Wobec tego, numery kolejnych interakcji: $1,2,\dots,l_I$ tworzą zbiór $L_I = \{1,2,\dots,l_I\}$. Co oczywiste: $|L_I| = l_I$.

Zbiór M to zbiór czasów kolejnych żądań (rozpoczęcia interakcji): $M = \{m_1, m_2, \dots, m_{l_I}\}$.

Na potrzeby dalszej części rozprawy wprowadźmy definicję żądania:

Definicja 4.1.2. *Żądanie e_l jest uporządkowaną 3-elementową krotką $(a_i\ u_k, m_l)$, gdzie: $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana usługa, $m_l \in M$ – czas pojawienia się żądania o numerze l.*

[19] Uporządkowanie żądań usług jedynie pod względem czasu mogłoby okazać się niewystarczające z uwagi chociażby na ograniczoną w rzeczywistych środowiskach rozdzielczość zegara (spójność ścisła w praktyce nie jest implementowalna w rozproszonym systemie wieloagentowym).

[20] Definicję interakcji zaprezentowano w kolejnym punkcie (4.1.3).

[21] Ze względu na to, że proces „nabywania" usługi, zaczyna się już w momencie identyfikacji potrzeby skorzystania z niej. W praktyce, w ramach modelu przyjmujemy, że czas od momentu identyfikacji potrzeby skorzystania z usługi do momentu wysłania żądania nie ma znaczenia.

[22] Dzięki sekwencyjnemu uszeregowaniu żądań, ta właściwość jest zagwarantowana. Przy zakładanym zerowym czasie trwania interakcji nie istnieje jednoczesność i zagwarantowany jest globalny porządek zdarzeń (zapewniony przez czas i identyfikator agenta). Co więcej, ze względu na to, że interakcje są natychmiastowe, porządek czasowy liniowy jest tożsamy z porządkiem sekwencyjnym.

Uwaga 4.1.2.1. $m_l \in M$ – to zarówno czas (moment) pojawienia się żądania o numerze l, jak i czas rozpoczęcia interakcji oraz czas zakończenia interakcji (na mocy powyższego założenia o pomijalnym czasie trwania interakcji). Z tego względu pojęcia te są stosowane zamiennie.

Uwaga 4.1.2.2. Zbiorem wszystkich żądań jest zbiór E, $E = \{e_1, e_2, \dots, e_{l_l}\}$, $|E| = l_I$

4.1.3. Interakcja i wynik interakcji

Jak określono wcześniej, Q jest zbiorem możliwych jakości usług. Zgodnie z uwagą 2.1. w rozdziale 2, interakcja obejmuje żądanie usługi przez pewnego agenta, oraz jej wyświadczenie na rzecz innego agenta. Formalnie funkcję interakcji[23] można zdefiniować w następujący sposób:

Definicja 4.1.3.1. *Funkcją interakcji jest funkcja częściowa* $f_{int}: A \times U \times M \times A \rightarrow Q$, *przy czym jeżeli* $f_{int}(a_i, u_k, m_l, a_j) = q^l$, *to* $a_i \in A$ *– usługobiorca (agent żądający usługi),* $u_k \in U$ *– żądana (świadczona) usługa,* $m_l \in M$ *– czas rozpoczęcia interakcji o numerze l,* $a_j \in A$ *– usługodawca (agent dostarczający usługę), a* $q^l \in Q$ *– jakość dostarczonej usługi w ramach* l*-tej interakcji (wynik tej interakcji).*

Uwaga 4.1.3.1. Interakcja i świadczenie usługi odbywa się na podstawie żądania skierowanego od usługodawcy do usługobiorcy, dlatego funkcja interakcji może być także określona przez funkcję częściową $f_{intE}: E \times A \rightarrow Q$, przy czym jeżeli $f_{intE}(e_l, a_j) = q^l$ to $e_l \in E$ – żądanie o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę), a $q^l \in Q$ – jakość dostarczonej usługi w ramach l-tej interakcji (wynik tej interakcji).

Warto zauważyć, że agent usługodawca jest parametrem funkcji interakcji, co oznacza, że zamiast zdefiniować jedną funkcję interakcji dla całego środowiska, każdy z agentów – usługodawców może zdefiniować własną funkcję interakcji. Te dwa ujęcia są w praktyce jednoznaczne, a przyjęto, że funkcja interakcji jest globalna dla całego środowiska, ze względu na to, że założenia twórców środowiska zwykle narzucają pewną postać funkcji interakcji, mimo tego, że każdy z agentów nadal ma możliwość własnej implementacji.

[23] Funkcją interakcji jest nazwana funkcja częściowa, gdyż potraktowano określenie „funkcja interakcji" jako nazwę własną. Dla większej precyzji można stosować nazwę „funkcja częściowa interakcji", ale zdecydowano się na pominięcie określenia „częściowa" dla większej przejrzystości. Podobny zabieg dokonano w dalszej części rozprawy, m.in. w odniesieniu do funkcji rekomendacji.

Definicja 4.1.3.2. *Zbiorem interakcji I jest dziedzina funkcji interakcji*[24].

Wyjaśnienie 4.1.3. Dziedziną funkcji częściowej jest zbiór wartości dla których funkcja częściowa jest zdefiniowana.

Uwaga 4.1.3.2. W różnych chwilach czasowych ta sama para agentów może wymieniać się tą samą usługą z różną jakością.

Definicja 4.1.3.3. *Interakcja $i_l \in I$ to element zbioru interakcji o numerze l (argument funkcji interakcji o numerze l), czyli krotka: (a_i, u_k, m_l, a_j), przy czym $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana (świadczona) usługa, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę).*

Definicja 4.1.3.4. *Ciąg wyników interakcji to ciąg $Q_{RES}: L_I \to Q$, gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru Q i odpowiadają kolejnym wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji.*

Uwaga 4.1.3.3. q^l oznacza element ciągu Q_{RES} o numerze l, czyli wynik interakcji o numerze l.

Warto zauważyć, że jakość usługi świadczonej przez usługodawcę w ogólnym przypadku nie musi być tożsama z jakością usługi dostarczonej do usługobiorcy, z tego względu wprowadźmy pojęcie funkcji zakłóceń:

Definicja 4.1.3.5. *Funkcją zakłóceń jest funkcja częściowa $f_{dis}: Z \times M \to O$, przy czym jeżeli $f_{dis}(z_l, m_l) = o^l$, to: $z_l \in Z$ – zdarzenie elementarne polegające na wystąpieniu zakłócenia o pewnej wartości wpływu na wynik interakcji, $m_l \in M$ – czas interakcji* o numerze l, *a $o^l \in O$ – jakość usługi w ocenie usługobiorcy dostarczonej w ramach interakcji (rzeczywisty wynik interakcji).*

[24] *funkcji częściowej $f_{int}: A \times A \times U \times M \to Q$ lub funkcji częściowej $f_{intE}: E \times A \to Q$*

Uwaga 4.1.3.4. Zbiór możliwych wyników interakcji O jest tożsamy ze zbiorem możliwych jakości usług Q, tj. $O = Q$, a jakość dostarczonej usługi w ramach interakcji jest rzeczywistym wynikiem tej interakcji.

Uwaga 4.1.3.5. W odniesieniu do porównania definicji 4.1.3.1. z definicją 4.1.3.5. przyjęto, że wystąpienie zakłócenia zgodnie z funkcją zakłóceń może wpłynąć jedynie negatywnie na jakość usługi dostarczonej usługobiorcy (i ocenionej przez niego) czyli, że jakość usługi dostarczonej jest zawsze nie większa niż jakość usługi wyświadczonej przez usługodawcę, tzn. że $\forall k \in [1, \dots, l_l]: o^k \leq q^k$. Czyli zakłócenie ma charakter zaburzenia poprawnego odbioru usługi, które jest niezależne zarówno od usługodawcy jak i usługobiorcy[25,26].

Definicja 4.1.3.6. *Ciąg rzeczywistych wyników interakcji to ciąg $O_{RES}: L_I \rightarrow O$, gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru O i odpowiadają kolejnym rzeczywistym wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji.*

Uwaga 4.1.3.6. o^l oznacza element ciągu O_{RES} o numerze l, czyli rzeczywisty wynik interakcji o numerze l.

4.1.4. Wybór usługodawcy

Funkcja wyboru usługodawcy określa sposób w jaki spośród zbioru potencjalnych usługodawców świadczących usługę u_k zostanie wybrany agent, od którego usługa u_k zostanie zażądana (z którym zostanie nawiązana interakcja).

Definicja 4.1.4. *Niech $\mathbb{A}$ będzie rodziną skończoną wszystkich podzbiorów zbioru A ($\mathbb{A} = P(A)$, $\mathbb{A}$ jest zbiorem potęgowym zbioru A). Wtedy każdy z elementów rodziny $\mathbb{A}$ jest pewnym podzbiorem zbioru A i można go utożsamiać ze zbiorem potencjalnych usługodawców.* ***Funkcją wyboru usługodawcy*** *jest funkcja częściowa $f_{sel}: \mathbb{A}\ x\ M \rightarrow A$, która dla dowolnego $X \in \mathbb{A}$ spełnia warunek $f_{sel}(X, m) \in X$.*

[25] Możliwe byłoby także uwzględnienie zakłócenia o innym charakterze – związanego z nieprawidłową oceną jakości usługi przez usługobiorcę. Takie zakłócenie byłoby związane z charakterystyką konkretnego agenta, w przeciwieństwie do zakłóceń oznaczonych na mocy definicji 4.1.3.5. które są określone dla całego środowiska.
[26] Teoretycznie możliwe są też pomyłki na korzyść odbiorcy, w przypadku gdyby można było dopuścić sytuację, w której rzetelny agent nie musi świadczyć usługi z maksymalną możliwą jakością Przyjęto jednak, że wpływ takich zakłóceń na ocenę rzetelności usługobiorcy byłby pomijalny, dlatego zdecydowano się przyjąć ograniczenie dotyczące tego, że zakłócenia mogą wpłynąć jedynie negatywnie na jakość dostarczanej usługi.

Uwaga 4.1.4.1. Alternatywnie, warunek w definicji 4.1.4. może być także wyrażony poprzez warunek:
jeżeli $X \in \mathbb{A}$, $f_{sel}(X, m) = Y$, to $\forall y \in Y: y \in X$, czyli każdy z elementów wartości funkcji częściowej wyboru usługodawcy musi być także elementem argumentu tej funkcji częściowej.

Własność 4.1.4.*:* Niezależnie od postaci funkcji wyboru usługodawcy, zawsze $f_{sel}(\emptyset, m) = \emptyset$. Własność 4.1.4. jest zgodna z intuicją, ponieważ jeżeli nie istnieje żaden agent, który mógłby świadczyć daną usługę to nie może być nawiązana interakcja.

Uwaga 4.1.4.2.. Argumentem funkcji wyboru usługodawcy w momencie m, może być pewien podzbiór X zbioru agentów $A_{P:u_l}$, którzy są w stanie świadczyć żądaną usługę u_l. Podzbiór $X \subseteq A_{P:u_l}$ nie musi obejmować całego zbioru agentów, którzy są w stanie świadczyć tę usługę ze względu na brak możliwości żądania usługi od niektórych agentów, wynikający z braku połączenia w warstwie usługowej pomiędzy agentem żądającym usługi, a potencjalnym usługodawcą. Inaczej mówiąc, nie wszystkie agenty świadczące danę usługę będą w stanie ją wyświadczyć dla agenta chcącego ją uzyskać.

Przykład 4.1.4.
Niech funkcja wyboru usługodawcy będzie miała następującą postać:

$$f_{sel}(X, m_l) = \begin{cases} \emptyset, X = \emptyset \\ x_j, X = \{x_0, \dots, x_{k-1}\} \end{cases}, X \subseteq A$$

przy czym:

l – numer interakcji o czasie rozpoczęcia m_l,

$$k = |X|,$$

$$l \equiv j \bmod k, 0 \leq j < k$$

Czyli wynikiem jest agent (element x_j) wybrany spośród podzbioru agentów świadczących daną usługę, będącego argumentem funkcji częściowej f_{sel}.
Załóżmy, że w danym środowisku agenty a_1, a_2, a_3 świadczą pewną usługę u_1. Wobec tego $A_{P:u_1} = \{a_1, a_2, a_3\}$. Wtedy wynikiem funkcji wyboru usługodawcy w pewnej interakcji m_l, czyli $f_{sel}(A_{P:u_1}, m_l) = f_{sel}(\{a_1, a_2, a_3\}, m_l)$ może być jeden z trzech agentów: a_1, a_2 , lub a_3. To, który z nich to będzie, jest uzależnione od numeru interakcji, np. dla pierwszej i czwartej

interakcji (m_1, m_4) będzie to agent a_1, dla drugiej (m_2) będzie to agent a_2, dla trzeciej (m_3) będzie to agent a_3.

Uwaga 4.1.4.3. W definicji 4.1.4. założono, że agent w środowisku może wybrać jednego agenta, od którego zażąda świadczenia usługi. W wielu środowiskach jest to słuszne założenie, gdyż agenty nie będą miały możliwości kolektywnego (wspólnego) świadczenia danej usługi. Jednak w ogólnym przypadku może być konieczny wybór jednocześnie większej liczby usługodawców. Jest to szczególnie istotne w przypadku gdy do skorzystania z danej usługi potrzeba jednoczesnej współpracy kilku usługodawców (np. w przypadku przekazywania pakietu w sieci, gdzie pomiędzy dostawcą a odbiorcą istnieją ścieżki, na których znajduje się wiele agentów/węzłów pośredniczących). Aby uwzględnić możliwość świadczenia danej usługi kolektywnie przez wielu usługodawców, definicja 4.1.4. mogłaby zostać uogólniona do definicji 4.1.4':

Definicja 4.1.4'. *Niech $\overline{\overline{\mathbb{A}}}$ będzie zbiorem potęgowym zbioru $\mathbb{A}$ ($\overline{\overline{\mathbb{A}}} = P(\mathbb{A}) = P(P(A))$, czyli $\mathbb{A}$ jest zbiorem potęgowym zbioru A, a $\overline{\overline{\mathbb{A}}}$ jest zbiorem potęgowym zbioru $\mathbb{A}$). Wtedy każdy z elementów $\overline{\overline{\mathbb{A}}}$ jest pewnym podzbiorem zbioru $\mathbb{A}$. Istnieje podzbiór $\overline{\overline{\mathbb{A}'}} \subset \overline{\overline{\mathbb{A}}}$, który zawera te elementy zbioru $\overline{\overline{\mathbb{A}}}$, (czyli te zbiory agentów) które mogą kolektywnie świadczyć usługę. Zbiór $\overline{\overline{\mathbb{A}'}}$ jest więc zbiorem potencjalnych usługodawców.* ***Funkcją wyboru usługodawcy*** *jest funkcja częściowa $f_{sel}: \overline{\overline{\mathbb{A}'}}\ x\ M \to \mathbb{A}$, która dla dowolnego $X \in \overline{\overline{\mathbb{A}'}}$ spełnia warunek $f_{sel}(X, m) \in X$.*

Uwaga 4.1.4.4. Jak zauważono wcześniej, w praktyce najczęściej wybierany będzie jeden agent do świadczenia usług (lub ewentualnie brak agentów). Wtedy, przy wykorzystaniu definicji 4.1.4' zbiór wartości funkcji częściowej f_{sel} będzie rodziną jednoelementowych podzbiorów zbioru A wraz ze zbiorem pustym, czyli będzie stanowił podzbiór rodziny $\mathbb{A}$.

Uwaga 4.1.4.5. Warto podkreślić, że argumentami funkcji wyboru usługodawcy są zbiory potencjalnych usługodawców. Inaczej mówiąc, istnieje ograniczona swoboda wyboru usługodawców, tzn. nie każdy podzbiór usługodawców jest poprawnym wyborem (np. nie tworzy ścieżki do celu w przypadku problemu routingu). Jak najbardziej jest też możliwa sytuacja, w której zbiór usługodawców świadczących potrzebną usługę nie jest pusty, ale rodzina podzbiorów możliwych do wybrania pusta jednak jest, bo żadna kombinacja nie spełnia warunków koniecznych wyboru.

W dalszej części rozprawy, aby niepotrzebnie nie komplikować modelu będziemy wykorzystywać definicję 4.1.4., ale warto podkreślić, że rozszerzenie modelu tak aby uwzględniał możliwość świadczenia usługi przez wielu usługodawców jest możliwe z wykorzystaniem definicji 4.1.4'.

4.1.5. Specyfikacja środowiska

Na podstawie wprowadzonych definicji i oznaczeń, możemy określić nową, formalną definicję środowiska:

Definicja 4.1.5. *Środowisko jest uporządkowaną 5-elementową krotką (5-ką)* (A, U, E, F, M)*, gdzie A jest zbiorem agentów (łącznie z ich charakterystyką), U jest zbiorem usług, E jest zbiorem żądań, F jest zbiorem funkcji częściowych określonych w środowisku* $F = \{f_{sel}, f_{int}, f_{dis}\}$*, a M jest zbiorem czasów interakcji.*

W celu jednoznacznego scharakteryzowania środowiska muszą zostać określone:

- agenty A działające w środowisku;
- usługi U świadczone w środowisku i zbiór możliwych jakości usług Q;
- dla każdego agenta a_k, usługi świadczone przez tego agenta: U_{a_k} – jest to składowa charakterystyki agentów;
- dla każdej usługi u_l świadczonej przez każdego z agentów a_k, maksymalna możliwa jakość tej usługi: $q_{a_k}^{u_l}$ – jest to składowa charakterystyki agentów;
- topologia agentów w warstwie usługowej (wskazująca, które agenty mogą nawiązywać ze sobą interakcje) – jest to składowa charakterystyki agentów;
- zbiór M czasów interakcji;
- zbiór żądań E;
- funkcja wyboru usługodawcy: f_{sel};
- funkcja interakcji: f_{int} lub f_{intE};
- funkcja zakłóceń: f_{dis}.

Uwaga 4.1.5. W środowisku, dla każdego żądania określonego w E, następuje w kolejności:

- wybór usługodawcy – zgodnie z funkcją f_{sel}.:

- interakcja – zgodnie z funkcją: f_{int} lub f_{intE};
- wystąpienie zakłócenia – zgodnie z funkcją: f_{dis};

Po czym następuje obsługa kolejnego żądania.

4.1.6. Przykład

Rozważmy środowisko będące platformą e-handlu, składające się z 20 agentów: $A = \{a_1, a_2, \dots, a_{20}\}$, w którym są świadczone dwie usługi (oferowane są dwa produkty na sprzedaż): $U = \{u_1, u_2\}$. Agenty o numerach od 1 do 5 są usługodawcami (ale mogą także korzystać z usług innych agentów), przy czym usługa u_1 jest świadczona przez agenty o numerach od 1 do 3, czyli: $A_{P:u_l} = \{a_1, a_2, a_3\}$, a usługa u_2 przez agenty 1,4 i 5, czyli: $A_{P:u_2} = \{a_1, a_4, a_5\}$. Wobec tego: $U_{a_1} = \{u_1, u_2\}$, $\forall_{k\epsilon\{2,3\}}\, U_{a_k} = \{u_1\}$, $\forall_{k\epsilon\{4,5\}}\, U_{a_k} = \{u_2\}$, $\forall_{k\epsilon\mathbb{N},5<k\leq 20}\, U_{a_k} = \{\emptyset\}$. Każdy agent może żądać świadczenia usługi od każdego z agentów świadczących daną usługę (mamy do czynienia z połączeniami każdego agenta z każdym agentem-usługodawcą na poziomie warstwy usługowej). W przypadku braku systemu TRM wybór usługodawcy odbywa się na zasadzie losowania z jednakowym prawdopodobieństwem spośród zbioru usługodawców, przy czym dla każdej interakcji funkcja wyboru usługodawcy f_{sel} jest ściśle określona. Usługi w środowisku mogą być świadczone z jakością $Q = \langle -1,1\rangle$, gdzie $q = 0$ odpowiada brakowi usługi, a ujemne wartości odpowiadają takiej jakości usługi, która wyrządza szkodę przy korzystaniu z niej. Każdy z usługodawców jest w stanie świadczyć usługę z maksymalną jakością $q = 1$, czyli: $\forall_{l\epsilon\{1,2\}}\, q_{a_1}^{u_l} = 1$, $\forall_{k\epsilon\{2,3\}}\, q_{a_k}^{u_1} = 1$, $\forall_{k\epsilon\{4,5\}}\, q_{a_k}^{u_2} = 1$. W przypadku gdy żaden z agentów nie stosuje ataku, każdy z agentów świadczy usługi z maksymalną możliwą jakością, $q = 1$, czyli: $\forall_{l,k}\, f_{intE}(e_l, a_k) = 1$. W środowisku nie występują zakłócenia, czyli: $\forall_l\, f_{dis}(z_l, m_l) = q^l$, a więc: $\forall_l\, o^l = q^l$.

Do przeprowadzenia analizy środowiska konieczne jest ustalenie sekwencji żądań usług. Przyjmijmy, że każdy z agentów, niebędący usługodawcą może wygenerować żądanie świadczenia usługi $u_1$lub u_2 (losując z której usługi skorzysta). Odbywa się to w taki sposób, że spośród agentów jest losowany agent, który wygeneruje żądanie, a następnie jakiej usługi zażąda. Warto zauważyć, że sposób generowania żądań nie jest częścią modelu, został przedstawiony jedynie dla ustalenia uwagi, natomiast model zakłada, że żądania dla danego środowiska są określone.

Dla tak określonego środowiska na bazie przedstawionego modelu można wykonywać dalsze analizy oraz zdefiniować system zarządzania zaufaniem. Jest to oczywiście środowisko bardzo uproszczone, zawierające niewiele agentów i niewiele usług, ale zostało

zaprezentowane jedynie w celu zobrazowania specyfiki modelu. W badaniach przeprowadzonych w rozdziale 6 przykład ten zostanie rozwinięty.

4.1.7. Ograniczenia modelu

Mimo tego, że przedstawiony model środowiska oddaje, zdaniem autora, w wystarczający sposób charakterystykę istotnych cech środowiska, to, jak każdy model, nie jest w stanie uwzględnić wszystkich aspektów rzeczywistego środowiska. W poniższych punktach opisano pokrótce ograniczenia modelu.

4.1.7.1. Homogeniczność a heterogeniczność usług

W praktyce, poszczególne typy usług, mimo tego że bardzo zbliżone, to mogą jednak się różnić pewnymi cechami. Dla przykładu: jeden ze sprzedających może sprzedawać pewien produkt w platformie e-commerce, a inny może oferować dokładnie taki sam produkt ale z krótszym czasem dostawy. W takim przypadku możliwe byłoby podejście, w którym są to odmienne usługi, ale mocno skorelowane ze sobą. Innym przykładem jest to, że oferowane są dwa produkty, które są bardzo do siebie podobne, ale różnią się pewną szczególną, nieistotną cechą. Warto zauważyć, że w celu skutecznego ujęcia takiej specyfiki, konieczne byłoby wprowadzenie dodatkowego pojęcia: substytutu lub podobieństwa usług. Wynika to z faktu, że mimo różnic przedstawionych w powyższych przykładach, usługi te z punktu widzenia usługobiorcy zaspokajają tę samą potrzebę. Wobec tego, w ramach niniejszego modelu, usługi z powyższych przykładów zostaną potraktowane jako takie same usługi, jest to uzasadnione tym, że dodatkowe cechy tych usług nie są znaczące. W takim podejściu istotne staje się rozstrzygnięcie, w którym momencie dane usługi różnią się na tyle, że powinny być opisane jako całkowicie odmienne. Regułą jaką warto przyjąć jest istotność cech danej usługi – jeżeli różnią się znacznie, czyli odpowiadają na inną potrzebę usługobiorcy, to warto je potraktować jako odmienne usługi (np. jeżeli w ramach platformy e-commerce są to różne produkty). Oczywiście jest to w pewien sposób subiektywne, a definicja usług zależy od dokonującego charakterystyki konkretnego środowiska, mimo to wydaje się, że w ogólnym przypadku to ograniczenie nie powinno wpływać na wyniki dokonywanej oceny wiarygodności systemów TRM.

4.1.7.2. Zmienność charakterystyki agentów

Pewnym uproszczeniem stosowanym w dalszych badaniach będzie założenie o tym, że agenty nie zmieniają swojej charakterystyki w czasie (np. przez cały rozważany okres są w stanie świadczyć te same usługi, czyli: U_{a_k} nie zależy od m_l oraz maksymalna jakość danej usługi także nie zmienia się w czasie, czyli: $q_{a_k}^{u_j}$ nie zależy od m_l). Co oczywiste, nie oznacza to, że agent a_k nie może w danej interakcji świadczyć usługi u_j o jakości niższej niż $q_{a_k}^{u_j}$ (czyli stosować atak) oraz że w różnych interakcjach nie może świadczyć usług o różnej jakości, oznacza to jedynie, że cały czas agent jest w stanie świadczyć te same usługi z taką samą maksymalną jakością. To ograniczenie ułatwia praktyczne zastosowanie modelu i dlatego warto je było przyjąć, natomiast istnieje możliwość łatwego rozszerzenia modelu w taki sposób aby był w stanie uwzględnić zmienną w czasie charakterystykę agentów – usługodawców. Aby tego dokonać wystarczyłoby wprowadzić zmienność zbioru agentów, tzn. przedział czasu, w którym dany agent jest aktywny. Zmianę jakiejkolwiek cechy agenta można wtedy zamodelować dekomponując go na kilka wirtualnych agentów aktywnych w różnych okresach.

4.1.7.3. Pomijalny czas trwania interakcji

Założenie o pomijalnym czasie trwania interakcji jest istotnym ograniczeniem modelu. W praktyce, w przypadku większości rzeczywistych środowisk, czas od momentu wysłania żądania, poprzez wyświadczenie usługi przez usługodawcę, aż do odbioru usługi przez usługobiorcę, jest niezerowy. Przyjęcie takiego ograniczenia umożliwia łatwiejszą późniejszą analizę, jak i istotnie zmniejsza skomplikowanie samego modelu. Konsekwencją takiego założenia jest fakt, że nigdy nie wystąpi sytuacja w której jednocześnie odbywa się więcej niż jedna interakcja. Ten fakt znacząco ułatwia analizę, ale z drugiej strony pewne strategiczne działania agentów nie będą mogły zostać wzięte pod uwagę. W szczególności nie będzie możliwe zamodelowanie przypadku, w którym pewien agent usługodawca otrzymuje wiele żądań świadczenia usług, których nie dostarcza, natomiast usługobiorcy nawet nie orientują się, że zostali oszukani, ponieważ czas dostarczenia usługi jeszcze nie upłynął.

Możliwe jest rozszerzenie modelu, tak aby uwzględnić czas trwania interakcji. W takim przypadku czas trwania interakcji (od momentu żądania do odbioru usługi) stałby się parametrem samej interakcji. W dalszym ciągu interakcje można by uszeregować względem czasu wysłania żądania danej usługi (co przestałoby być tożsame z momentem interakcji, gdyż ta byłaby rozłożona w czasie) i identyfikatora agenta żądającego. Autor zdecydował się na pominięcie tego aspektu z uwagi na to, że nawet jego nieuwzględnienie umożliwi

przeprowadzenie wartościowych analiz (co zostanie przedstawione w dalszej części rozprawy), a jednocześnie pozwoli uniknąć nadmiernego ich skomplikowania. Jednocześnie kwestię uwzględnienia czasu interakcji należy traktować jako interesujące przyszłe zagadnienie badawcze.

4.2. MODEL SYSTEMU TRM

Agenty funkcjonujące w ramach środowiska i wykorzystujące system zarządzania zaufaniem i reputacją, mogą wymieniać się na temat innych agentów rekomendacjami, które mogą zostać wykorzystane do oceny zaufania do innych agentów lub reputacji innego agenta. Na podstawie oceny zaufania lub reputacji, agent żądający danej usługi może dokonać wyboru agenta usługodawcy, z którym wejdzie w interakcję (proces świadczenia usługi). Po interakcji możliwa jest aktualizacja wartości zaufania lub reputacji do agentów istniejących w systemie (lub podzbioru agentów).

Zbiór możliwych działań agenta obejmuje co najmniej następujące działania (odpowiadające etapom funkcjonowania systemu TRM, określonym w podrozdziale 2.5):

1. Pozyskanie informacji, w tym żądanie rekomendacji
2. Ocena zaufania lub ocena reputacji
3. Wybór agenta świadczącego usługę
4. Interakcja (świadczenie usługi) i ocena jej wyniku
5. Aktualizacja wartości zaufania lub reputacji

4.2.1. Zaufanie lub reputacja

Częściowo właściwości zaufania i reputacji omówiono w podrozdziale 2.4, w niniejszym punkcie przedstawiono formalny opis zaufania i reputacji. Zarówno zaufanie, jak i reputacja dotyczą określonego kontekstu (zdefiniowanego w punkcie 4.2.2).

Definicja 4.2.1.1. *T jest zbiorem możliwych wartości zaufania określonych przez system TRM.*

Uwaga 4.2.1.1. Wartość zaufania $t_n \in T$ może być wyrażona przez pojedynczą wartość lub przez wektor/krotkę (np. zawierającą wartość będącą oszacowaniem zaufania oraz drugą wartość – charakteryzującą pewność tego oszacowania). Dalsze rozważania są przeprowadzone w odniesieniu do zaufania jako pojedynczej wartości, a w przypadku zaufania wyrażonego jako

wektor należy, zależnie od jej interpretacji w danym TRM, dookreślić odpowiednie zależności, np. relację porządku.

Uwaga 4.2.1.2. Wartość zaufania $t_{min} \in T$ to minimalna wartość zaufania, taka, że $\forall t_n \in T$: $t_{min} \leq t_n$, a wartość zaufania $t_{max} \in T$ to maksymalna wartość zaufania, taka, że $\forall t_n \in T$: $t_{max} \geq t_n$.

Definicja 4.2.1.2. *Zaufaniem jest funkcja częściowa* $f_{trust}: A \times A \times C \times M \to T$, *przy czym jeżeli* $f_{trust}(a_i, a_j, c_k, m_l) = t_n$, *to* $a_i \in A$ *- agent ufający,* $a_j \in A$ *- agent zaufany,* $c_k \in C$ *– kontekst zaufania,* $m_l \in M$ *– czas (interakcji lub oceny zaufania),* $t_n \in T$ *– wartość zaufania.*

Zarówno agent a_i jak i a_j (lub jeden z nich) w powyższej definicji mogą być zastąpione pewną grupą agentów. Wtedy mamy do czynienia z zaufaniem grupowym, a definicja 4.2.1.2. może przyjąć postać definicji 4.2.1.2'. W praktyce, zwykle jest jednak stosowane zaufanie pomiędzy konkretną parą agentów, dlatego w dalszym ciągu będziemy stosować definicję 4.2.1.2.

Definicja 4.2.1.2'. *Niech* $\mathbb{A}$ *będzie zbiorem potęgowym zbioru A bez zbioru pustego (*$\mathbb{A} = P(A) - \{\emptyset\}$*). Wtedy każdy z elementów rodziny* $\mathbb{A}$ *jest pewnym podzbiorem zbioru A i można go utożsamiać z pewną grupą agentów. Zaufanie może być określone jako funkcja częściowa* $f_{trust}: \mathbb{A} \times \mathbb{A} \times C \times M \to T$, *przy czym jeżeli* $f_{trust}(A_i, A_j, c_k, m_l) = t_n$, *to* $A_i \in \mathbb{A}$ *– zbiór agentów ufających,* $A_j \in \mathbb{A}$ *– zbiór agentów zaufanych,* $c_k \in C$ *– kontekst zaufania,* $m_l \in M$ *– czas (interakcji lub oceny zaufania),* $t_n \in T$ *– wartość zaufania.*

Na potrzeby dalszej części rozprawy, wprowadźmy następujące oznaczenia dotyczące wartości zaufania w chwili m_l:

- $t_{a_i \to a_j}^{c_k; m_l}$ – wartość zaufania agenta a_i do agenta a_j w kontekście c_k w chwili m_l;
- $\vec{t}_{a_i \to a_j}^{C; m_l}$ – wektor wartości zaufania agenta a_i do agenta a_j we wszystkich kontekstach w chwili m_l;
- $T_{a_i \to A}^{C; m_l}$ – macierz wartości zaufania agenta a_i do wszystkich agentów we wszystkich kontekstach w chwili m_l;

- $\vec{t}_{a_i \to A}^{c_k;m_l}$ – wektor wartości zaufania agenta a_i do wszystkich agentów w kontekście c_k w chwili m_l;
- $T_{A \to A}^{c_k;m_l}$ – macierz wartości zaufania wszystkich agentów do wszystkich agentów w kontekście c_k w chwili m_l.

Reputacja jest pojęciem analogicznym do zaufania z tym, że jest ona charakterystyką pewnego agenta w środowisku, a nie charakterystyką relacji pomiędzy parą agentów (lub grupami agentów wedle definicji 4.2.1.2'.).

Definicja 4.2.1.3. *P jest zbiorem możliwych wartości reputacji określonych przez system TRM.*

Uwaga 4.2.1.3. Wartość reputacji $p_n \in P$ może być wyrażona przez pojedynczą wartość lub przez wektor/krotkę (np. zawierającą wartość będącą oszacowaniem reputacji oraz drugą wartość – charakteryzującą pewność tego oszacowania). Dalsze rozważania są przeprowadzone w odniesieniu do reputacji jako pojedynczej wartości, a w przypadku reputacji wyrażonej jako wektor należy, zależnie od jej interpretacji w danym TRM, dookreślić odpowiednie zależności, np. relację porządku.

Uwaga 4.2.1.4. Wartość reputacji $p_{min} \in P$ to minimalna wartość reputacji, taka, że $\forall p_n \in P$: $p_{min} \leq p_n$, a wartość reputacji $p_{max} \in P$ to maksymalna wartość reputacji, taka, że $\forall p_n \in P$: $p_{max} \geq p_n$.

Definicja 4.2.1.4. *Reputacją jest funkcja częściowa* $f_{rep}: A \times C \times M \to P$*, przy czym jeżeli* $f_{rep}(a_j, c_k, m_l) = p_n$*, to* $a_j \in A$ *- agent obdarzony reputacją (zaufany),* $c_k \in C$ *– kontekst reputacji,* $m_l \in M$ *– czas (interakcji lub oceny reputacji),* $p_n \in P$ *– wartość reputacji.*

Uwaga 4.2.1.5. Możliwe jest traktowanie reputacji jako szczególnego przypadku zaufania w myśl definicji 4.2.1.2'., w którym grupa agentów ufających jest zbiorem A wszystkich agentów w środowisku (czyli $A_i = A$).

Podobnie jak w przypadku zaufania, reputacja także może charakteryzować grupę agentów, a nie pojedynczego agenta, wtedy definicja 4.2.1.4. może przyjąć postać definicji

4.2.1.4'. W praktyce, zwykle jest jednak stosowana reputacja charakteryzująca konkretnego agenta, dlatego w dalszej części rozprawy będziemy stosować definicję 4.2.1.4.

Definicja 4.2.1.4'. *Niech* $\mathbb{A}$ *będzie zbiorem potęgowym zbioru* A *bez zbioru pustego (*$\mathbb{A} = P(A) - \{\emptyset\}$*). Wtedy każdy z elementów rodziny* $\mathbb{A}$ *jest pewnym podzbiorem zbioru* A *i można go utożsamiać z pewną grupą agentów. Reputacja może być określona jako funkcja częściowa* $f_{rep}: \mathbb{A} \times C \times M \to P$*, przy czym jeżeli* $f_{rep}(A_j, c_k, m_l) = p_n$*, to* $A_j \in \mathbb{A}$ *– zbiór agentów obdarzonych reputacją (zaufanych),* $c_k \in C$ *– kontekst reputacji,* $m_l \in M$ *– czas (interakcji lub oceny reputacji),* $p_n \in P$ *– wartość reputacji.*

Analogicznie jak w przypadku zaufania możemy wprowadzić następujące oznaczenia dotyczące wartości reputacji w chwili m_l:

- $p_{a_j}^{c_k;m_l}$ – wartość reputacji agenta a_j w kontekście c_k w chwili m_l;
- $\vec{p}_{a_j}^{C;m_l}$ – wektor wartości reputacji agenta a_j we wszystkich kontekstach w chwili m_l;
- $\vec{p}_{A}^{c_k;m_l}$ – wektor wartości reputacji wszystkich agentów w kontekście c_k w chwili m_l;
- $P_{A}^{C;m_l}$ – macierz wartości reputacji wszystkich agentów we wszystkich kontekstach w chwili m_l.

Warto zauważyć, że zwykle w systemie TRM jest zdefiniowane zaufanie lub reputacja, przy czym rzadko są używane obydwie te funkcje jednocześnie.

4.2.2. Kontekst

Tak jak w życiu społecznym rzadko mamy do czynienia z zaufaniem absolutnym (w każdym możliwym aspekcie) lub absolutną rekomendacją (tzn. taką, która dotyczy wszystkiego), tak i w systemach TRM zwykle zaufanie, reputacja lub rekomendacja posiada określony kontekst. W najogólniejszym ujęciu, kontekstem jest to czego dotyczy zaufanie, reputacja lub rekomendacja. Może istnieć dowolna liczba kontekstów, zaufanie w każdym kontekście jest interpretowane tak jak określa to system TRM.

W dalszym rozważaniu skupimy się na aspekcie zaufania, jednak należy pamiętać, że takie samo rozumowanie można przeprowadzić dla reputacji i rekomendacji[27]. Najczęściej wykorzystywanym kontekstem jest świadczenie usługi (dowolnej), a zaufanie w takim

[27] Definicja rekomendacji znajduje się w punkcie 4.2.3.

kontekście dotyczy przewidywań co do tego, że usługi zostaną wyświadczone na oczekiwanym poziomie. W przeciwieństwie do tak zdefiniowanego ogólnego kontekstu – świadczenia wszystkich usług w danym środowisku, możliwe powinno być zdefiniowanie oddzielnych kontekstów dla każdej z usług, lub dla pewnego podzbioru usług świadczonych w środowisku. W takim przypadku zaufanie dotyczące świadczenia usługi u_1 (a dokładniej: zaufanie do agenta, będące miarą wiary w rzetelność świadczenia tej usługi przez tego agenta), może być zupełnie inne niż zaufanie do tego samego agenta dotyczące świadczenia usługi u_2. Jest to zgodne z intuicją, np. w relacjach społecznych można mieć wysokie zaufanie, że pewien dziennikarz będzie tworzył rzetelne artykuły, ale jednocześnie zaufanie, że będzie on w stanie świadczyć usługi medyczne, powinno być niskie. W takim ujęciu kontekst jest określony poprzez podzbiór zbioru usług $U_k \subseteq U$ i dotyczy rzetelności świadczenia tych usług. Taki poziom kontekstu (świadczenia usług) określmy jako poziom zerowy zagnieżdżenia kontekstu.

Niemniej, kontekst nie musi być związany tylko ze świadczeniem usług, ale może dotyczyć rekomendacji dotyczącej świadczenia usług (z określonego zbioru). Skoro możliwe jest otrzymanie rekomendacji w określonym kontekście to powinno być możliwe wyznaczenie zaufania do tej rekomendacji (czyli stworzenie kolejnego poziomu zagnieżdżenia kontekstu). Taki kontekst może być wykorzystany do oceny zaufania do rekomendacji dotyczącej świadczenia pewnych usług. Analogicznie, poszukując lekarza możemy się kierować rekomendacjami na temat różnych lekarzy, ale zdecydowanie większe zaufanie będziemy mieli do rekomendacji dobrego i uczciwego znajomego, który z usług tego lekarza skorzystał, niż do rekomendacji znajomego, który jedynie słyszał o tym lekarzu, a tym bardziej do rekomendacji całkowicie anonimowej. W takim ujęciu kontekst jest określony przez rekomendację dotyczącą świadczenia usług z określonego zbioru (ozn. $r(U_k)$). Taki poziom kontekstu (rekomendacji dotyczącej świadczenia usług) określmy jako poziom pierwszy zagnieżdżenia kontekstu.

Stosując analogiczne rozumowanie możemy wyznaczyć także zaufanie do rekomendacji dotyczącej rekomendacji dotyczącej świadczenia usług definiując kolejny kontekst (ozn. $r(r(U_k))$). Taki poziom kontekstu (rekomendacji dotyczącej rekomendacji dotyczącej świadczenia usług) określamy jako poziom drugi zagnieżdżenia kontekstu.

Powyższe rozumowanie można powtarzać w nieskończoność, ale w praktyce sens jest rozważać jedynie kontekst do pewnego poziomu zagnieżdżenia.

Określony kontekst może także składać się z pewnego zbioru wcześniej zdefiniowanych kontekstów, np. mając zaufanie do mechanika samochodowego, możemy mieć także zaufanie do jego rekomendacji dotyczących lakiernika samochodowego.

Powyższe rozważania umożliwiają przedstawienie formalnej definicji kontekstu.

Definicja 4.2.2.1. *Kontekst $c_k \in C_\alpha$ określa to czego dotyczy rekomendacja lub miara zaufania lub reputacji.*

Kontekstem jest:

- *świadczenie usług z dowolnego niepustego podzbioru zbioru wszystkich usług: $c_k = U_k$, gdzie $U_k \subseteq U$, $U_k \neq \emptyset$, $U = \{u_1, \dots, u_l\}$; lub*
- *rekomendacja dotycząca dowolnego kontekstu: $c_l = r(c_k)$; lub*
- *dowolny zbiór kontekstów: $c_m = \{c_k, c_l, \dots\}$.*

Uwaga 4.2.2.1. W ogólnym przypadku, zbiór wszystkich możliwych kontekstów C_α, na podstawie powyższej definicji, jest nieskończony, gdyż dla każdego istniejącego kontekstu c_k można utworzyć np. kontekst $c_l = r(c_k)$. Niemniej, w każdym systemie TRM musi być zdefiniowany skończony zbiór kontekstów $C \subset C_\alpha$ wykorzystywanych w systemie.

Przykład 4.2.2.1. Załóżmy, że w środowisku świadczone są usługi u_1 i u_2, czyli: $U = \{u_1, u_2\}$. Wtedy możliwe są następujące konteksty:

- konteksty świadczenia usług (zerowy poziom zagnieżdżenia kontekstu): $c_1 = \{u_1\}$, $c_2 = \{u_2\}$, $c_3 = \{u_1, u_2\}$;
- konteksty rekomendacji dotyczących świadczonych usług (pierwszy poziom zagnieżdżenia kontekstu): $c_4 = r(\{u_1\})$, $c_5 = r(\{u_2\})$, $c_6 = r(\{u_1, u_2\})$;
- konteksty rekomendacji dotyczących rekomendacji dotyczących świadczonych usług: $c_7 = r\big(r(\{u_1\})\big)$, $c_8 = r\big(r(\{u_2\})\big)$, $c_9 = r\big(r(\{u_1, u_2\})\big)$;
- konteksty na kolejnych poziomach zagnieżdżenia kontekstu, np. $c_{10} = r\Big(r\big(r(\{u_1\})\big)\Big)$, $c_{13} = r\Big(r\big(r\big(r(\{u_1\})\big)\big)\Big)$, $c_{16} = r\Big(r\Big(r\big(r\big(r(\{u_1\})\big)\big)\Big)\Big)$, itd.
- konteksty zbiorowe, np. $c_k = \Big\{\{u_1\}, r(\{u_1\}), r\big(r(\{u_1\})\big), \dots\Big\}$ – dotyczy świadczenia usługi u_1, jak i wszelkich rekomendacji (na dowolnym poziomie zagnieżdżenia kontekstu) na temat świadczenia usługi u_1.

W rzeczywistych systemach TRM, nie są wykorzystywane wszystkie możliwe konteksty, wedle powyższej definicji, dlatego konieczne jest zdefiniowanie zbioru kontekstów, które są określone w danym systemie TRM:

Definicja 4.2.2.2. *Zbiór kontekstów C zawiera konteksty, dla których mogą być określone wartości zaufania lub reputacji lub dla których mogą być wydawane rekomendacje w danym systemie TRM. Zbiór kontekstów C jest skończonym podzbiorem zbioru wszystkich możliwych kontekstów: $C \subset C_\alpha$.*

W praktyce często systemy TRM definiują następujące konteksty:

Przykład 4.2.2.2.1. W wielu systemach TRM, w szczególności takich, które nie wykorzystują rekomendacji, jest wykorzystywany jedynie kontekst świadczenia usług (wszystkich), tzn.: $C = \{c_1\}$, $c_1 = U$ – wtedy wszelkie wartości zaufania (lub reputacji) dotyczą rzetelności świadczenia usług (wszystkich łącznie).

Przykład 4.2.2.2.2. W przypadku systemów TRM wykorzystujących rekomendacje często wykorzystywane są dwa konteksty $C = \{c_1, c_2\}$, przy czym:

- $c_1 = U$ – kontekst świadczenia usług (wszystkich), zaufanie w tym kontekście dotyczy rzetelności świadczenia usług; w przypadku niektórych systemów zaufanie w tym kontekście jest nazywane zaufaniem akcyjnym lub zaufaniem usługowym;
- $c_2 = r(U)$ – kontekst rekomendacji dotyczących świadczonych usług, zaufanie w tym kontekście dotyczy rzetelności wydawanych rekomendacji; w przypadku niektórych systemów zaufanie w tym kontekście jest nazywane zaufaniem rekomendacyjnym lub zaufaniem do rekomendacji.

Przykład 4.2.2.2.3. System RefTRM [45] definiuje następujące trzy konteksty $C = \{c_1, c_2, c_3\}$ w odniesieniu do zaufania:

- $c_1 = U$ – kontekst świadczenia usług (wszystkich), zaufanie w tym kontekście jest nazywane zaufaniem akcyjnym;
- $c_2 = r(U)$ – kontekst rekomendacji dotyczących świadczonych usług, zaufanie w tym kontekście jest nazywane zaufaniem rekomendacyjnym;
- $c_3 = \{U, r(U)\}$ – kontekst zaufania zbiorowego, zaufanie w tym kontekście jest nazywane w ramach systemu RefTRM zaufaniem całkowitym.

Przykład 4.2.2.2.4. Mimo istnienia rekomendacji możliwe jest istnienie tylko jednego kontekstu $C = \{c_1\}$, $c_1 = \{U, r(U)\}$ – wtedy występuje jedna miara zaufania do danego agenta zarówno w odniesieniu do świadczonych przez niego usług jak i wydawanych rekomendacji.

Przykład 4.2.2.2.5. Możliwa jest także sytuacja, w której definiowane są oddzielne konteksty w odniesieniu do świadczenia każdej usługi, oraz dodatkowy kontekst w odniesieniu do

rekomendacji dotyczących świadczenia usług (wszystkich), wtedy gdy: $U = \{u_1, \dots, u_l\}$, to $C = \{c_1, \dots, c_l, c_{l+1}\}$, przy czym:

- $\forall_{k:1 \le k \le l}\ c_k = \{u_k\}$ – kontekst świadczenia usługi u_k,
- $c_{l+1} = r(U)$ – kontekst wydawanych rekomendacji.

Uwaga 4.2.2.2. Możliwe jest rozgraniczenie zbioru kontekstów zaufania lub reputacji od zbioru kontekstów rekomendacji. Wtedy w odniesieniu do zaufania (lub reputacji) mogą być stosowane inne konteksty niż w odniesieniu do rekomendacji. Takie rozgraniczenie miałoby sens zwłaszcza o ile zbiór kontekstów rekomendacji zawierałby się w zbiorze kontekstów zaufania. Wtedy zbiór kontekstów zaufania mógłby zawierać dodatkowy kontekst będący kontekstem zaufania do rekomendacji najwyższego poziomu zagnieżdżenia kontekstu. Takie rozgraniczenie nie spowodowałoby znacznego skomplikowania modelu, ale mimo tego w dalszej części pracy nie będzie stosowane. Wobec tego przez zbiór kontekstów będziemy rozumieć zarówno zbiór kontekstów zaufania (lub reputacji) jak i zbiór kontekstów rekomendacji.

4.2.3. Rekomendacje i ich rodzaje

W wielu systemach TRM poszczególne agenty nie tylko oceniają zaufanie do innych agentów lub reputację innych agentów, ale także wymieniają się rekomendacjami na temat innych agentów lub na temat interakcji, które zaszły w przeszłości. Postać rekomendacji i zbiór możliwych wartości rekomendacji zależy od konkretnego systemu TRM. W szczególności rekomendacją może być wartość zaufania wedle własnej oceny przekazana innemu agentowi (i tak się dzieje w przypadku niektórych systemów), ale w ogólnym przypadku wartość rekomendacji może być określona oddzielnie od zaufania, choć z całą pewnością główną ideą systemów TRM jest to aby wydawane rekomendacje na temat pewnego agenta były zależne od wartości zaufania do tego agenta (lub oceny jakości interakcji z tym agentem).

Rekomendacja może mieć formę własnej oceny zaufania lub prostej wskazówki (np. nie korzystaj), albo wręcz własnych wyników interakcji.

Rekomendacja może dotyczyć oceny rzetelności agenta (będziemy ją nazywać rekomendacją agenta) lub może dotyczyć opinii o konkretnej interakcji (będziemy ją nazywać rekomendacją interakcji). Poniżej dokonano określenia formalnej definicji poszczególnych typów rekomendacji.

Definicja 4.2.3.1. *R jest zbiorem możliwych wartości rekomendacji określonych przez system TRM.*

Uwaga 4.2.3.1. Wartość rekomendacji $r_n \in R$ może być wyrażona przez pojedynczą wartość lub przez wektor/krotkę (np. zawierającą wartość będącą rekomendacją oraz drugą wartość – charakteryzującą pewność rekomendacji – w ocenie samego agenta). Dalsze rozważania są przeprowadzone w odniesieniu do rekomendacji jako pojedynczej wartości, a w przypadku rekomendacji wyrażonej jako wektor należy, zależnie od jej interpretacji w danym TRM, dookreślić odpowiednie zależności, np. relację porządku.

4.2.3.1. Rekomendacja agenta

Definicja 4.2.3.2. *Funkcją rekomendacji wewnętrznej agenta jest funkcja częściowa* $f_{reca\_in}: A \times A \times C \times M \rightarrow R$, *przy czym jeżeli* $f_{reca\_in}(a_j, a_p, c_k, m_l) = r_n$, *to* $a_j \in A$ – *agent dostarczający rekomendację (wydawca rekomendacji),* $a_p \in A$ – *agent, którego dotyczy rekomendacja (przedmiot rekomendacji),* $c_k \in C$ – *kontekst rekomendacji,* $m_l \in M$ – *czas,* $r_n \in R$ – *wartość rekomendacji wewnętrznej agenta.*

Definicja 4.2.3.3. *Rekomendacją wewnętrzną agenta jest krotka 5-elementowa* $(a_j, a_p, c_k, m_l, r_n)$, *utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji wewnętrznej agenta o jej wartość.*

Definicja 4.2.3.4. *Funkcją rekomendacji agenta jest funkcja częściowa* $f_{reca}: A \times A \times A \times C \times M \rightarrow R$, *przy czym jeżeli* $f_{reca}(a_i, a_j, a_p, c_k, m_l) = r_n$, *to* $a_i \in A$ – *agent żądający rekomendacji (odbiorca rekomendacji),* $a_j \in A$ – *agent dostarczający rekomendację (wydawca rekomendacji),* $a_p \in A$ – *agent, którego dotyczy rekomendacja (przedmiot rekomendacji),* $c_k \in C$ – *kontekst rekomendacji,* $m_l \in M$ – *czas,* $r_n \in R$ – *wartość rekomendacji.*

Definicja 4.2.3.5. *Rekomendacją agenta jest krotka 6-elementowa* $(a_i, a_j, a_p, c_k, m_l, r_n)$, *utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji agenta o jej wartość. Zbiór takich krotek 6-elementowych jest zbiorem rekomendacji agenta* R^{AS}.

4.2.3.2. Rekomendacja interakcji

Definicja 4.2.3.6. *Funkcją rekomendacji wewnętrznej interakcji jest funkcja częściowa* $f_{reci\_in}: A \times I \times C \times M \rightarrow R$, *przy czym jeżeli* $f_{reci\_in}(a_j, i_p, c_k, m_l) = r_n$, *to* $a_j \in A$ *– agent dostarczający rekomendację (wydawca rekomendacji),* $i_p \in I$ *– element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji),* $c_k \in C$ *– kontekst rekomendacji,* $m_l \in M$ *– czas,* $r_n \in R$ *– wartość rekomendacji.*

Definicja 4.2.3.7. *Rekomendacją wewnętrzną interakcji jest krotka 5-elementowa* $(a_j, i_p, c_k, m_l, r_n)$, *utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji wewnętrznej interakcji o jej wartość.*

Definicja 4.2.3.8. *Funkcją rekomendacji interakcji jest funkcja częściowa* $f_{reci}: A \times A \times I \times C \times M \rightarrow R$, *przy czym jeżeli* $f_{reci}(a_i, a_j, i_p, c_k, m_l) = r_n$, *to* $a_i \in A$ *– agent żądający rekomendacji (odbiorca rekomendacji),* $a_j \in A$ *– agent dostarczający rekomendację (wydawca rekomendacji),* $i_p \in I$ *– element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji),* $c_k \in C$ *– kontekst rekomendacji,* $m_l \in M$ *– czas,* $r_n \in R$ *– wartość rekomendacji.*

Definicja 4.2.3.9. *Rekomendacją interakcji jest krotka 6-elementowa* $(a_i, a_j, i_p, c_k, m_l, r_n)$, *utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji interakcji o jej wartość. Zbiór takich krotek 6-elementowych jest zbiorem rekomendacji interakcji* R^{IS}.

4.2.3.3. Rekomendacja

Termin 4.2.3.1. *Rekomendacja to rekomendacja agenta lub rekomendacja interakcji.*

Termin 4.2.3.2. *Zbiorem rekomendacji* R^S *jest suma zbioru rekomendacji agenta i zbioru rekomendacji interakcji,* $R^S = R^{AS} \cap R^{IS}$.

4.2.3.4. Pozyskiwanie i przechowywanie rekomendacji

Rekomendacje mogą być pozyskiwane w różnych momentach (np. przed lub po interakcji), mogą także być wydawane w odpowiedzi na żądanie innych agentów lub w określonej sytuacji. Jak zauważono wcześniej, istotne jest także kto może mieć dostęp do

danej rekomendacji oraz gdzie i jak długo jest ona przechowywana i użyteczna. Niniejszy punkt przedstawia najistotniejsze podtypy rekomendacji w różnych aspektach. Dany system TRM może stosować określone podtypy rekomendacji.

Rozważając sposób żądania i czas wydania, rekomendacje mogą być wydawane:

- bezpośrednio po interakcji (zwykle wtedy są to rekomendacje interakcji);
- na żądanie agenta (wydawane przed wybraniem usługodawcy w ramach danej interakcji, zwykle wtedy są to rekomendacje agenta).

W zależności od odbiorcy, rekomendacja może być:

- dedykowana dla agenta (wtedy ta konkretna rekomendacja jest niedostępna dla innych agentów, ale zwykle inne agenty mogą pozyskać własne dedykowane rekomendacje);
- dedykowana dla podzbioru agentów;
- publiczna (dostępna dla wszystkich agentów w środowisku)[28];
- niedostępna dla agentów (dostępna jedynie dla centralnego podsystemu zarządzania rekomendacjami – RMSS)[29].

W aspekcie czasu życia, rekomendacja może być:

- ulotna (wykorzystywana natychmiast przez odbiorcę, a następnie usuwana);
- o określonym czasie ważności;
- trwała (dostępna do końca funkcjonowania odbiorcy w środowisku).

W aspekcie miejsca przechowywania, rekomendacja może być:

- nieprzechowywana (dotyczy to zwykle rekomendacji ulotnej) – nie istnieje takie miejsce, ponieważ rekomendacja nie jest przechowywana;
- przechowywana lokalnie (w pamięci agenta);
- przechowywana centralnie (w pewnej pamięci dostępnej w ramach środowiska, np. w ramach podsystemu zarządzania rekomendacjami – RMSS).

W aspekcie identyfikowalności wydawcy[30], rekomendacja może być:

- identyfikowalna (odbiorca rekomendacji posiada wiedzę, który agent wydał rekomendację);
- pseudoanonimowa (odbiorca rekomendacji nie zna identyfikatora wydawcy lub zna pewien identyfikator wydawcy rekomendacji ale nie jest w stanie powiązać tego

[28] Dyskusja dotycząca konsekwencji istnienia publicznych rekomendacji została przedstawiona w punkcie 4.2.3.7.
[29] Podsystem zarządzania rekomendacjami (RMSS) został omówiony w punkcie 4.2.3.6.
[30] Aspekt ten jest szerzej dyskutowany w punkcie 4.2.3.7.

identyfikatora z konkretnym agentem w środowisku, natomiast podsystem zarządzania rekomendacjami jest w stanie zidentyfikować wydawcę rekomendacji);

- anonimowa (nie można powiązać rekomendacji z jej wydawcą).

Na podstawie powyższej klasyfikacji można określić cechy rekomendacji używanych w danym systemie TRM, przy czym należy pamiętać że podstawowa klasyfikacja obejmuje to, czy rekomendacja dotyczy agenta, czy interakcji.

Przykład 4.2.3.1. W popularnej platformie e-handlu allegro.pl, rekomendacje używane w systemie TRM to rekomendacje interakcji, wydawane po interakcji, częściowo publiczne i częściowo niedostępne dla agentów[31], trwałe, centralnie przechowywane oraz identyfikowalne.

4.2.3.5. Żądanie i wybór dostawców rekomendacji

System TRM powinien określać sposób wyboru agentów, którzy wydadzą rekomendacje. W szczególności mogą być używane wszystkie rekomendacje istniejące w systemie. Poniżej zdefiniowana funkcja wyboru dostawcy rekomendacji ma zastosowanie jedynie w systemach TRM, w których rekomendacje są wydawane na żądanie agenta. W przypadku systemów TRM używających rekomendacji wydawanych bezpośrednio po interakcji, wykorzystywane są te rekomendacje, do których agent ma dostęp lub pewne agregaty takich rekomendacji (np. średnia wartość rekomendacji). Funkcja wyboru dostawców rekomendacji ma postać:

Definicja 4.2.3.10. *Niech* $\mathbb{A}$ *będzie zbiorem potęgowym zbioru* A ($\mathbb{A} = P(A)$). *Wtedy każdy z elementów* $\mathbb{A}$ *jest pewnym podzbiorem zbioru* A *i można go utożsamiać ze zbiorem potencjalnych dostawców rekomendacji.* ***Funkcją wyboru dostawców rekomendacji*** *jest funkcja częściowa* $f_{sel\_rec}: C \times A \times M \to \mathbb{A}$, *przy czym jeżeli* $f_{sel\_rec}(c_i, a_j, m_l) = A_p$, *to* $c_i \in C$ *– kontekst rekomendacji,* $a_k \in A$ *– podmiot rekomendacji,,* $m_l \in M$ *– czas,* $A_p \in \mathbb{A}$ *– zbiór dostawców rekomendacji.*

Do każdego agenta ze zbioru dostawców rekomendacji wysyłane jest żądanie rekomendacji.

[31] W platformie allegro niepełna rekomendacja jest publiczna, ponieważ oprócz podstawowej informacji (rekomendacja pozytywna lub negatywna) i komentarza słownego, kupujący ocenia sprzedającego dodatkowo w wybranych aspektach (w skali pięciostopniowej) i ta część nie jest dostępna dla żadnego kupującego (agenta), ale jest używana do stworzenia przez centralny system zagregowanej oceny sprzedającego.

Definicja 4.2.3.11. *Żądanie rekomendacji $e_{R:l}$ jest uporządkowaną 4-elementową krotką (a_i, a_p c_k, m_l) lub (a_i, i_p c_k, m_l), gdzie: $a_i \in A$ – żądający rekomendacji, $a_p \in A$ – przedmiot rekomendacji (agent) lub $i_p \in I$ – przedmiot rekomendacji (interakcja), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas pojawienia się żądania o numerze l.*

4.2.3.6. Podsystem zarządzania rekomendacjami (RMSS)

Rekomendacje mogą być przesyłane bezpośrednio pomiędzy agentami, ale także agenty mogą wykorzystywać w tym celu podsystem RMSS (ang. „Recommendation Management Subsystem") – jeżeli został on uwzględniony w ramach danego systemu TRM. Podsystem RMSS pełni funkcję pośrednika przy przekazywaniu informacji (żądań wydania rekomendacji i wydanych rekomendacji), a także może pełnić funkcję agregatora lub repozytorium wydanych rekomendacji.

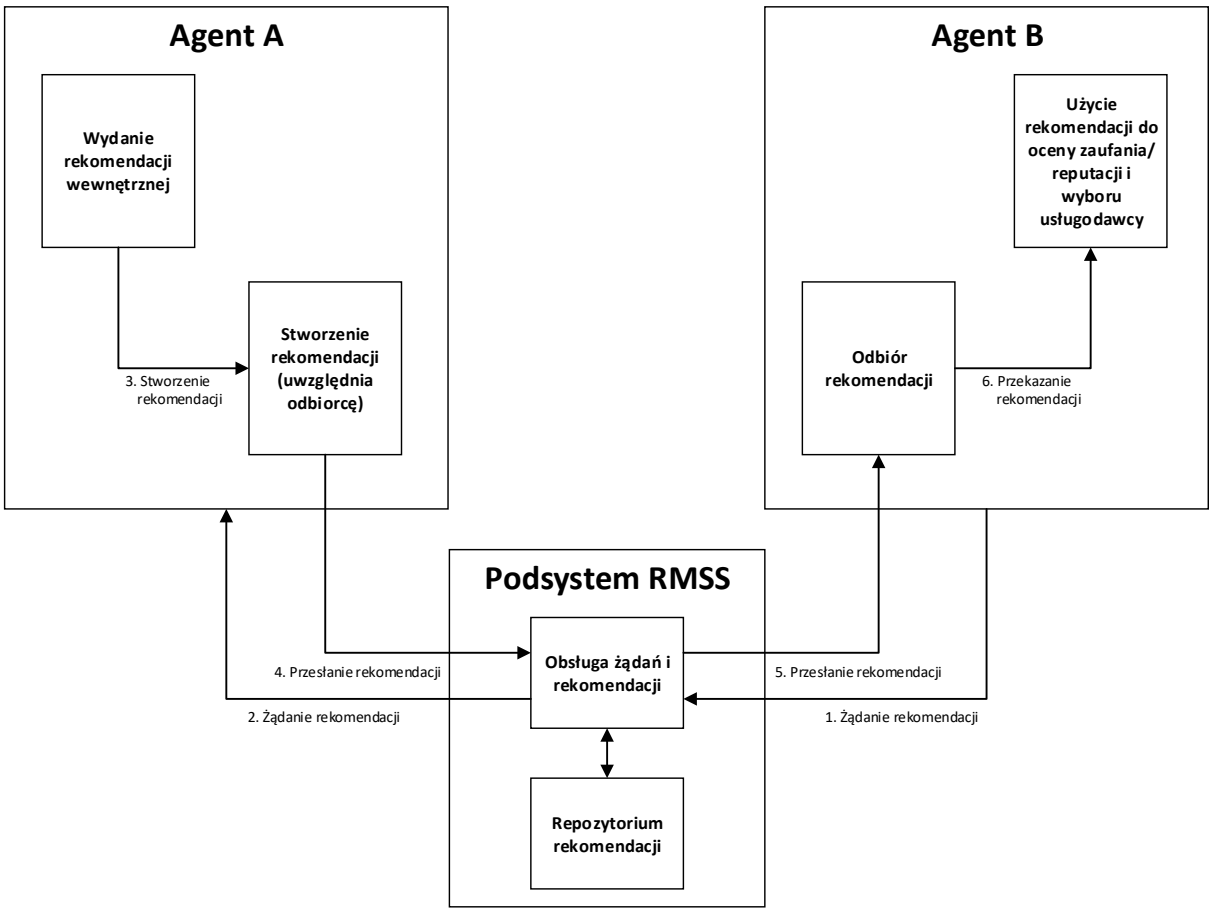
Podsystem RMSS może także być wykorzystywany do tłumaczenia wartości rekomendacji, np. w sytuacji gdy jeden z agentów stosuje inny system TRM (i inny zbiór wartości rekomendacji), niż drugi agent[32]. Co więcej, wykorzystanie podsystemu RMSS pozwala także na buforowanie i przechowywanie rekomendacji – jest to przydatne w sytuacji gdy używane są różne TRM, a w jednym z nich rekomendacje są wydawane zawsze po skorzystaniu z usługi i nie mogą zostać wydane w innym momencie, podczas gdy w innym systemie TRM żądania wydania rekomendacji napływają przed skorzystaniem z usługi. Wykorzystanie podsystemu RMSS pozwala więc na współpracę rożnych, teoretycznie niekompatybilnych systemów TRM używanych przez poszczególne agenty. Oczywiście podsystem RMSS musi w takim przypadku „rozumieć" wszystkie wykorzystywane systemy TRM, w szczególności znać ich zbiory wartości rekomendacji.

Przekazywanie rekomendacji pomiędzy parą agentów przedstawiono schematycznie na rysunku 8.

Osobny podsystem RMSS nie jest oczywiście niezbędny do funkcjonowania systemu TRM – możliwa i jak najbardziej spotykana jest konstrukcja systemu, w której agenty przekazują sobie rekomendacje bezpośrednio. Dla ujednolicenia modelu sytuacja taka jest modelowana poprzez wprowadzenie wirtualnego, trywialnego RMSS, który jedynie zapewnia mechanizm przekazywania żądań i odpowiedzi – innymi słowy sama sieć komunikacyjna

[32] Współpraca różnych TRM w ramach jednego środowiska jest interesującym zagadnieniem badawczym, prawie nieeksplorowanym w literaturze. Zagadnienie to wykracza poza zakres niniejszej rozprawy, niemniej jest potencjalnym obszarem przyszłych prac.

łącząca agenty może być traktowana jako trywialny rodzaj podsystemu RMSS. W przypadku systemów TRM, w których występuje centralne miejsce, gdzie są przechowywane rekomendacje agentów, rolę taką pełni właśnie podsystem RMSS i w takim przypadku nie dochodzi do bezpośredniej komunikacji agentów ze sobą, jeżeli chodzi o przekazywanie komunikatów systemu TRM. Warto zauważyć, że wprowadzenie takiego podsystemu umożliwia także to, aby część komunikacji TRM była realizowana w praktyce bezpośrednio pomiędzy agentami (wtedy RMSS jedynie przekazuje komunikaty), a część zadań była realizowana w RMSS. Dzięki temu możliwe jest osiągnięcie wystarczającej ogólności i możliwości modelowania znacznej liczby różnych typów systemów TRM.



Rysunek 8 Przekazywanie rekomendacji

4.2.3.7. Rekomendacja publiczna i anonimowa

Niniejszy punkt prezentuje kilka uwag dotyczących rekomendacji publicznych oraz anonimowych, gdyż zastosowanie rekomendacji tego typu ma istotny wpływ na możliwość zastosowania i skuteczność niektórych ataków na systemy TRM.

Uwaga 4.2.3.1. Rekomendacja może być przeznaczona dla pewnej grupy agentów[33] lub wręcz być dostępna dla wszystkich agentów (czyli stać się publiczna). Wtedy w definicjach 4.2.3.4. i 4.2.3.8. należy zastąpić agenta żądającego rekomendacji przez zbiór agentów, a w przypadku rekomendacji publicznej przyjąć, że zbiór agentów żądających rekomendacji jest tożsamy ze zbiorem wszystkich agentów.

Uwaga 4.2.3.2. Można sobie wyobrazić konstrukcję systemu TRM, w której wydawca rekomendacji nie jest znany odbiorcy rekomendacji (wydawca może być zanonimizowany). Taka sytuacja nie wpływa na kształt powyższych definicji, ale ma istotne konsekwencje w perspektywie ataków.

Uwaga 4.2.3.3. Jeżeli istnieje przypadek taki, że dla tego samego agenta dostarczającego rekomendację i agenta, którego dotyczy rekomendacja (lub interakcji, której dotyczy rekomendacja) istnieje taka para agentów żądających rekomendacji, że wartość rekomendacji jest różna, to mamy do czynienia z dyskryminacją ze względu na żądającego rekomendacji (może to być identyfikator ataku).

Uwaga 4.2.3.4. Stosowanie rekomendacji anonimowej lub pseudoanonimowej uniemożliwia obliczanie wartości zaufania rekomendacyjnego do wydawcy rekomendacji przez odbiorcę rekomendacji (samodzielnie, bez wykorzystania podsystemu RMSS). W przypadku rekomendacji pseudoanonimowej możliwe jest natomiast obliczanie zaufania do samej rekomendacji na podstawie prywatnego (znanego tylko zaufanej trzeciej stronie – taką rolę może pełnić podsystem RMSS) zaufania lub reputacji.

Jeżeli rekomendacje są publiczne lub wydawca rekomendacji jest anonimowy (lub pseudoanonimowy) to taki fakt ma istotne konsekwencje w perspektywie:

- poziomu wiedzy agentów,
- możliwości wydawania różnych rekomendacji dla różnych agentów,
- możliwości dokonywania analizy dynamicznej (zmian rekomendacji dotyczących tego samego agenta w czasie).

[33] Analogicznie można wyobrazić sobie sytuację, w której wystawca rekomendacji jest pewną grupą agentów, ale w praktyce autor nie spotkał się z taką konstrukcją systemu TRM.

Uwaga 4.2.3.5. Anonimizacja wydawcy rekomendacji lub publiczny charakter rekomendacji mogą uniemożliwiać część ataków. Na przykład w serwisie allegro.pl rekomendacje są publiczne, natomiast nie ma anonimizacji wydawcy rekomendacji (rekomendacja jest identyfikowalna). W takim przypadku wydawca rekomendacji nie ma możliwości stosowania dyskryminacji (przekazywania do różnych agentów różnych rekomendacji dotyczących agenta lub interakcji).

4.2.3.8. Zbiory dostępnych i wydanych rekomendacji

Definicja 4.2.3.12. *Zbiorem rekomendacji wydanych w systemie TRM do chwili m_z jest zbiór $R^{S:m_z}$, który zawiera wszystkie rekomendacje wydane w systemie TRM do chwili m_z (dla których $m_l \leq m_z$, $m_l \in M$ – czas); zbiór $R^{S:m_z}$ jest podzbiorem zbioru rekomendacji R^S, tj. $R^{S:m_z} \subseteq R^S$.*

Na potrzeby dalszej części rozprawy, wprowadźmy następujące oznaczenia dotyczące wartości rekomendacji:

- $r_{a_i \to a_j : a_p}^{c_k ; m_l}$ – wartość rekomendacji wydanej przez agenta a_i przesłanej do agenta a_j na temat agenta a_p w kontekście c_k w chwili m_l[34];
- $r_{a_i \to a_j : i_p}^{c_k ; m_l}$ – wartość rekomendacji wydanej przez agenta a_i przesłanej do agenta a_j na temat interakcji i_p w kontekście c_k w chwili m_l;
- $r_{a_i \to A : a_p}^{c_k ; m_l}$ – wartość rekomendacji publicznej (dostępnej dla wszystkich agentów), wydanej przez agenta a_i na temat agenta a_p w kontekście c_k w chwili m_l;
- $r_{a_i \to A : i_p}^{c_k ; m_l}$ – wartość rekomendacji publicznej (dostępnej dla wszystkich agentów), wydanej przez agenta a_i na temat interakcji i_p w kontekście c_k w chwili m_l;

4.2.3.9. Wydawanie rekomendacji

Moment wydawania oraz odbioru rekomendacji są określone przez typ rekomendacji. Ogólnie rzecz ujmując, wydawane rekomendacje mogą brać pod uwagę:

- wartości zaufania lub reputacji;

[34] W podobny sposób można zdefiniować oznaczenia dotyczące wartości rekomendacji wydawanej dla zbioru agentów.

- wyniki historycznych interakcji;
- inne rekomendacje.

Dokładny sposób tworzenia wydawanej rekomendacji, określenia jej wartości, a co za tym idzie postać funkcji rekomendacji, jest właściwością konkretnego systemu TRM.

4.2.3.10. Ocena rekomendacji

W przypadku niektórych systemów TRM otrzymane rekomendacje mogą podlegać ocenie, np. w oparciu o rzeczywiste interakcje, wykonane po otrzymaniu danej rekomendacji.

Definicja 4.2.3.13. *Algorytm oceny rekomendacji jest sposobem postępowania w celu aktualizacji wartości zaufania do agenta, który wydał rekomendację (lub reputacji tego agenta).*

W ogólnym przypadku algorytm oceny rekomendacji może uwzględniać następujące czynniki:

- wydaną rekomendację przez danego agenta (a także wcześniejsze rekomendacje agenta, jeżeli są dostępne);
- wartości zaufania lub reputacji (do agenta wydającego rekomendację, lub agenta będącego podmiotem rekomendacji);
- inne rekomendacje;
- wynik interakcji z agentem, którego dotyczyła rekomendacja.

4.2.4. Obserwacja i doświadczenie

W przypadku niektórych systemów TRM, agenty mogą obserwować interakcje innych agentów. Obserwacje mogą być wykorzystywane do oceny zaufania.

Definicja 4.2.4. *Obserwacją jest funkcja częściowa* $f_{obs}: A \times A \times A \times U \times M \rightarrow O$*, przy czym jeżeli* $f_{obs}(a_i, a_j, a_k, u_l, m_n) = o^n$*, to* $a_i \in A$ *– agent obserwujący,* $a_j \in A$ *– agent żądający usługi,* $a_k \in A$ *– agent dostarczający usługę,* $u_l \in U$ *– usługa,* $m_n \in M$ *– czas,* $o^n \in O$ *– rzeczywisty wynik tej interakcji pomiędzy agentami* a_j *i* a_k *zaobserwowany przez agenta* a_i*.*

Uwaga 4.2.4.1. Założono, dla ustalenia uwagi, że agent obserwujący zaobserwuje rzeczywisty wynik interakcji, ale w ogólnym przypadku można wyobrazić sobie sytuację, że dokona on obserwacji wyniku interakcji (przed wystąpieniem ewentualnego zakłócenia), albo wskutek własnej zaburzonej percepcji zaobserwuje zniekształcony wynik interakcji.

Uwaga 4.2.4.2. Doświadczeniem jest obserwacja w której agent obserwujący jest agentem żądającym usługi lub agentem dostarczającym usługę.

4.2.5. Czas pozyskania rekomendacji oraz oceny i aktualizacji wartości zaufania lub reputacji

Zgodnie z przedstawionym modelem, do podjęcia decyzji dotyczącej wyboru usługodawcy (zgodnie z określoną funkcją częściową wyboru usługodawcy) i nawiązania z tym usługodawcą interakcji, potrzebne jest pozyskanie rekomendacji i wykorzystanie oceny zaufania. Wobec tego pozyskanie rekomendacji zachodzi wcześniej niż sama interakcja. Podobnie po interakcji, może zachodzić aktualizacja oceny zaufania lub rekomendacji na podstawie wyniku tej interakcji. Natomiast zgodnie z przyjętym modelem wszystkie te działania zachodzą w czasie $m_i \in M$, czyli zakładamy, że każde z tych działań ma pomijalny czas trwania. W rzeczywistych systemach jednak nie będzie to prawda, z tym że w większości przypadków założenie o zerowym czasie poszczególnych działań nie powinno wpływać na funkcjonowanie środowiska. Warto zauważyć, że możemy zrezygnować z założenia o pomijalnym czasie trwania działań takich jak pozyskanie rekomendacji czy aktualizacja wartości zaufania i reputacji, a będzie to szczególnie łatwe o ile założymy, że wszystkie działania związane z poszczególną interakcją zakończą się przed jakimkolwiek działaniem związanym z kolejną interakcją.

4.2.6. Aktualizacja zaufania i reputacji

Jak określono wcześniej, zaufanie i reputacja to funkcja częściowa, ale istotny jest sposób obliczania (aktualizacji) wartości zaufania i reputacji jako wyniku tej funkcji. W ogólnym przypadku do poprawnego określenia systemu TRM konieczne jest wskazanie:

- dla których agentów i kontekstów wartość zaufania i reputacji będą wyznaczane;
- kiedy będzie następować aktualizacja tych wartości;
- jak będzie ona wykonywana (w jaki sposób obliczane będą nowe wartości).

Określać to powinien algorytm oceny interakcji i agenta:

Definicja 4.2.6. *Algorytm oceny interakcji i agenta jest sposobem postępowania w celu aktualizacji wartości zaufania do agenta, który uczestniczył w interakcji (lub reputacji tego agenta).*

Ogólnie rzecz ujmując, do obliczenia wartości zaufania lub reputacji mogą być brane pod uwagę:

- historyczne wartości zaufania;
- wyniki interakcji;
- obserwacje;
- rekomendacje.

Aktualizacja wartości zaufania lub reputacji może wynikać zarówno z funkcjonowania algorytmu oceny rekomendacji jak i algorytmu oceny interakcji i agenta.

4.2.7. Wpływ rekomendacji i zaufania lub reputacji na wybór usługodawcy i interakcje

Główną ideą funkcjonowania systemu zarządzania zaufaniem i reputacją jest to, żeby pozwalały one wpływać na wybór usługodawcy (lub ewentualnie wpływać na jakość świadczonych usług). W związku z tym, wykorzystywanie systemu TRM wpływa na zmianę postaci funkcji wyboru usługodawcy: f_{sel} oraz ewentualnie funkcji interakcji: f_{int} lub f_{intE} w taki sposób, że uwzględniają one dodatkowe charakterystyki agentów (nieistniejące w przypadku braku systemu TRM), takie jak reputacja agenta lub zaufanie do tego agenta. W związku z tym specyfikacja systemu TRM musi określać w jaki sposób miary zaufania i reputacji są wykorzystywane do podjęcia decyzji dotyczącej wyboru usługodawcy i ewentualnie w jaki sposób wpływają na jakość świadczonej usługi.

4.2.8. Specyfikacja systemu TRM

Na podstawie wprowadzonych definicji i oznaczeń, możemy określić nową, formalną definicję środowiska z działającym systemem TRM, które jest nieco rozbudowane w porównaniu do środowiska bez systemu TRM, określonego w definicji 4.1.5::

Definicja 4.2.8. *Środowisko z działającym systemem TRM jest uporządkowaną 9-elementową krotką (9-ką)* $(A', U, E, F', M, T, P, C, R)$*, gdzie* A' *jest zbiorem agentów (łącznie z ich charakterystyką),* U *jest zbiorem usług,* E *jest zbiorem żądań,* F' *jest zbiorem funkcji*

częściowych określonych w środowisku z systemem TRM, M jest zbiorem czasów interakcji, T jest zbiorem wartości zaufania, P jest zbiorem wartości reputacji, C jest zbiorem kontekstów, a R jest zbiorem rekomendacji.

Uwaga 4.2.8.1. Zbiór funkcji częściowych określonych w środowisku z systemem TRM może obejmować funkcje: f_{sel} , f_{int}, f_{dis}, f_{trust}, f_{rep}, f_{obs}, f_{rec}, f_{sel_rec} lub niektóre z nich.

Uwaga 4.2.8.2. W porównaniu do środowiska określonego w definicji 4.1.5., w środowisku z działającym systemem TRM charakterystyka agentów jest zmieniona (ponieważ może zawierać np. wartości zaufania do innych agentów), a w zbiorze funkcji częściowych występują dodatkowe funkcje f_{trust} lub f_{rep}, a także ewentualnie funkcje: f_{obs}, f_{rec} i f_{sel_rec}, ponadto funkcje f_{sel} , f_{int}, zależą od rozszerzonej charakterystyki agentów.

W celu jednoznacznego scharakteryzowania środowiska z działającym systemem TRM muszą zostać określone następujące elementy systemu TRM (jako dodatkowe w stosunku do elementów środowiska wymienionych w punkcie 4.1.5):

- zbiór możliwych wartości zaufania T lub zbiór możliwych wartości reputacji P;
- funkcje częściowe zaufania f_{trust} lub reputacji f_{rep} wraz z algorytmem oceny interakcji i agenta;
- zbiór używanych kontekstów C zaufania lub reputacji lub rekomendacji;
- jeżeli możliwe są obserwacje to funkcja częściowa obserwacji f_{obs};
- w przypadku istnienia rekomendacji:
    - typ używanej rekomendacji,
    - sposób funkcjonowania podsystemu zarządzania rekomendacjami RMSS,
    - zbiór możliwych wartości rekomendacji R,
    - funkcja częściowa rekomendacji f_{rec},
    - sposób generowania żądania rekomendacji $e_{R:l}$,
    - funkcja częściowa wyboru dostawców rekomendacji f_{sel_rec},
    - algorytm oceny rekomendacji;
- funkcja częściowa wyboru usługodawcy: f_{sel} z uwzględnieniem zaufania lub reputacji;
- funkcja interakcji: f_{int} lub f_{intE} z uwzględnieniem zaufania lub reputacji;
- wartości początkowe zaufania lub reputacji.

Uwaga 4.2.8.3. W środowisku, dla każdego żądania określonego w E, następuje w kolejności:

- pozyskanie rekomendacji (jeżeli są możliwe) – zgodnie z funkcją f_{sel_rec};
- ocena zaufania lub reputacji – zgodnie z funkcją f_{trust} lub f_{rep};
- wybór usługodawcy – zgodnie z funkcją f_{sel};
- interakcja – zgodnie z funkcją: f_{int} lub f_{intE};
- wystąpienie zakłócenia – zgodnie z funkcją: f_{dis};
- obserwacja (jeżeli jest możliwa w danym środowisku) – zgodnie z funkcją: f_{obs};
- ocena rekomendacji (jeżeli występuje);
- ocena lub aktualizacja zaufania lub reputacji (jeżeli występuje).

Po czym następuje obsługa kolejnego żądania.

Warto zauważyć, że w powyższym schemacie działania środowiska z systemem TRM, dwukrotnie występuje ocena zaufania lub reputacji, wynika to z faktu, że w różnych systemach TRM moment tej oceny może być różny, np. może występować tylko przed nawiązaniem interakcji lub także po interakcji.

4.2.9. Przykład

Rozważmy środowisko będące siecią sensorów, składające się z 20 agentów ($n = 20$ – liczba agentów w systemie): $A = \{a_1, a_2, \dots, a_{20}\}$, w którym świadczona jest jedna usługa: $U = \{u_1\}$. Agenty o numerach od 1 do 5 są usługodawcami, przy czym usługa u_1 jest świadczona przez agenty o numerach od 1 do 5, czyli: $A_{P:u_l} = \{a_1, a_2, a_3, a_4, a_5\}$. Wobec tego: $\forall_{k\in\mathbb{N},1\leq k\leq 5}\ U_{a_k} = \{u_1\}$, $\forall_{k\in\mathbb{N},5<k\leq 20}\ U_{a_k} = \{\emptyset\}$. Każdy agent nieświadczący usługi może żądać świadczenia usługi od każdego z agentów świadczących daną usługę (mamy do czynienia z połączeniami każdego agenta z każdym agentem-usługodawcą na poziomie warstwy usługowej). Usługi w środowisku mogą być świadczone z jakością $Q = \langle 0{,}1\rangle$, gdzie $q = 0$ odpowiada brakowi usługi, a $q = 1$ usłudze o najwyższej jakości. Każdy z usługodawców jest w stanie świadczyć usługę z maksymalną jakością $q = 1$, czyli: $\forall_{k\in\mathbb{N},1\leq k\leq 5}\ q_{a_k}^{u_1} = 1$. W przypadku gdy żaden z agentów nie stosuje ataku, każdy z agentów świadczy usługi z maksymalną możliwą jakością, $q = 1$, czyli: $\forall_{l,k}\ f_{intE}(e_l, a_k) = 1$. W środowisku nie występują zakłócenia, czyli: $\forall_l\ f_{dis}(z_l, m_l) = q^l$, a więc: $\forall_l\ o^l = q^l$.

Następnie zdefiniujmy system TRM działający w tym środowisku (definicja tego systemu została oparta o system RefTRM, przedstawiony w artykule [45], ale zastosowane oryginalnie oznaczenia zostały zmienione, tak aby dopasować je do przedstawionego modelu).

W systemie stosowane jest zaufanie, przy czym zbiór możliwych wartości zaufania $T = \langle 0,1 \rangle$, w systemie nie występuje reputacja (to pojęcie jest stosowane w artykule ale w odmiennym znaczeniu – w odniesieniu do pewnej zagregowanej miary wartości zaufania). W systemie używane są trzy konteksty zaufania: $C = \{c_1, c_2, c_3\}$:

- $c_1 = U$ – kontekst świadczenia usług, zaufanie w tym kontekście jest nazywane zaufaniem akcyjnym;
- $c_2 = r(U)$ – kontekst rekomendacji dotyczących świadczonych usług, zaufanie w tym kontekście jest nazywane zaufaniem rekomendacyjnym;
- $c_3 = \{U, r(U)\}$ – kontekst zaufania zbiorowego, zaufanie w tym kontekście jest nazywane w ramach systemu RefTRM zaufaniem całkowitym.

Funkcja f_{trust} jest zdefiniowana oddzielnie dla każdego z kontekstów i omówiona poniżej w punkcie 4.2.9.2.

Obserwacje nie są wykorzystywane.

Rekomendacje są używane i są to rekomendacje agenta, wydawane na żądanie agenta, dedykowane dla agenta, ulotne, nieprzechowywane i identyfikowalne. Nie istnieje jawnie zdefiniowany podsystem RMSS, z uwagi na to, że poszczególne rekomendacje są bezpośrednio przesyłane pomiędzy agentami. Zbiór możliwych wartości rekomendacji jest następujący: $R = \langle 0,1 \rangle$. Specyficznym aspektem tego systemu jest to, że przed każdą interakcją, agent-usługobiorca żąda rekomendacji nt. potencjalnych usługodawców od wszystkich pozostałych agentów działających w środowisku: $\forall_{a_k \in A, m_l \in M}\ f_{sel\_rec}(c_1, a_k, m_l) = A - a_k$; po czym do każdego z nich jest wysyłane żądanie rekomendacji ($e_{R:l}$). Rekomendacje są przesyłane jedynie w kontekście $c_1 = U$, czyli świadczenia usług. Wartość rekomendacji dotyczącej agenta a_k jest wartością zaufania akcyjnego (w kontekście c_1) do tego agenta, czyli: $\forall_{a_i, a_j, a_p \in A, m_l \in M}\ f_{rec} = f_{reca}(a_i, a_j, a_p, c_1, m_l) = t^{c_1; m_l}_{a_j \to a_p}$, czyli $\forall_{a_i, a_j, a_p \in A, m_l \in M}\ r^{c_1; m_l}_{a_i \to a_j : a_p} = t^{c_1; m_l}_{a_j \to a_p}$.

Funkcja częściowa wyboru usługodawcy f_{sel} zostanie omówiona poniżej w punkcie 4.2.9.3.

Funkcja interakcji f_{int} ma następującą postać $\forall_{a_j \in A_{P:u_1}, a_i \in A_R, m_l \in M}\ f_{int}(a_i, u_1, m_l, a_j) = q^{u_1}_{a_j}$ (oczywiście o ile agent a_j nie stosuje żadnego ataku), czyli nie zależy od zaufania (w żadnym kontekście), a jakość dostarczanej usługi jest równa maksymalnej jakości usługi jaką może świadczyć usługodawca. Ponieważ w środowisku nie występują zakłócenia to $\forall_l\ o_l = q_l$ (rzeczywisty wynik interakcji jest tożsamy z wynikiem interakcji).

Na początku funkcjonowania systemu poszczególne wartości zaufania w kontekście c_1 i c_2 są następujące:

$$\forall_{a_i,a_j \in A}\ t_{a_i \to a_j}^{c_1;m_0} = t_{init}^{c_1}$$

$$\forall_{a_i,a_j \in A}\ t_{a_i \to a_j}^{c_2;m_0} = t_{init}^{c_2}$$

Gdzie m_0 to czas początku działania środowiska (przed pierwszą interakcją), a $t_{init}^{c_1} \in \langle 0,1 \rangle$ to tzw. początkowe zaufanie akcyjne oraz a $t_{init}^{c_2} \in \langle 0,1 \rangle$ to tzw. początkowe zaufanie rekomendacyjne. W badaniach w przytoczonym artykule przyjęto, że: $t_{init}^{c_1} = t_{init}^{c_2} = 0.5$. Wartość zaufania w kontekście c_3 jest obliczana na podstawie wzoru w punkcie 4.2.9.2.

Na tej podstawie możemy przystąpić do opisania poszczególnych etapów funkcjonowania środowiska z systemem RefTRM, w odniesieniu do obsługi pojedynczego żądania.

4.2.9.1. Pozyskanie rekomendacji

Pozyskanie rekomendacji, zgodnie z funkcją f_{sel_rec} zostało omówione wcześniej – rekomendacje są pozyskiwane od wszystkich agentów na temat każdego potencjalnego usługodawcy.

4.2.9.2. Ocena zaufania lub reputacji

Ocena zaufania w kontekście c_3 (całkowitego) następuje na podstawie wartości zaufania w kontekście c_1 (akcyjnego) oraz otrzymanych rekomendacji i zaufania w kontekście c_2 (rekomendacyjnego) agentów dostarczających rekomendację, zgodnie z funkcją f_{trust};

$$f_{trust}(a_i, a_j, c_3, m_l) = t_{a_i \to a_j}^{c_3;m_l} = \alpha * t_{a_i \to a_j}^{c_1;m_l} + (1-\alpha) * \frac{\sum_{k=1}^{n} r_{a_k \to a_i:a_j}^{c_1;m_l} * t_{a_i \to a_k}^{c_2;m_l}}{\sum_{k=1}^{n} t_{a_i \to a_k}^{c_2;m_l}}$$

gdzie $\alpha \in (0,1)$ – parametr systemu TRM: tzw. waga zaufania akcyjnego.
Zaufanie w kontekście c_3 (zaufanie całkowite) agenta a_i do agenta a_j jest więc średnią ważoną zaufania agenta a_i do agenta a_j w kontekście c_1 (zaufania akcyjnego) oraz średniej wartości rekomendacji na temat agenta a_j pochodzącej od wszystkich innych agentów – średnia wartość rekomendacji jest ważona zaufaniem w kontekście c_2 (zaufaniem rekomendacyjnym) do agenta wydającego daną rekomendację.

4.2.9.3. Wybór usługodawcy

Wybór usługodawcy następuje na podstawie wartości zaufania w kontekście c_3 (całkowitego) i odbywa się na jeden z dwóch sposobów (określeń funkcji f_{sel}):

1. Wybierany jest agent z najwyższą wartością zaufania całkowitego[35]. O ile wartość tego zaufania przekracza pewien ustalony próg minimalnego zaufania: h
2. Agenty są szeregowane w kolejności malejącej wartości zaufania całkowitego, a następnie wybierany jest losowo z pewnym prawdopodobieństwem μ. W przypadku agentów, do których wartość zaufania całkowitego jest większa od ustalonego progu minimalnego zaufania: h, prawdopodobieństwo wyboru μ jest proporcjonalne do wartości tego zaufania, w przypadku agentów, do których wartość zaufania całkowitego jest mniejsza od ustalonego progu minimalnego zaufania, prawdopodobieństwo wyboru jest dodatkowo zmniejszane o β^S (parametr systemu – zmniejszenie prawdopodobieństwa wyboru na dostawcę usługi), wobec tego prawdopodobieństwo $\mu(a_j)$ wyboru agenta a_i jako dostawcę usługi dla agenta a_j jest proporcjonalne do następujących wartości:

$$\mu(a_i) \sim \begin{cases} t_{a_i \to a_j}^{c_3;m_{l-1}}, jeżeli: t_{a_i \to a_j}^{c_3;m_{l-1}} \geq h \\ \max\left(t_{a_i \to a_j}^{c_3;m_{l-1}} - \beta^S, 0\right), jeżeli: t_{a_i \to a_j}^{c_3;m_{l-1}} < h \end{cases}$$

4.2.9.4. Interakcja

Przebieg interakcji, został omówiony wcześniej – następuje świadczenie usług zawsze z maksymalną jakością[36] (niezależnie od zaufania do usługodawcy do usługobiorcy).

4.2.9.5. Wystąpienie zakłócenia

Wystąpienie zakłócenia, zgodnie z funkcją f_{dis}, zostało omówione wcześniej – zakłócenia nie występują.

4.2.9.6. Obserwacja

Obserwacja nie występuje (brak zdefiniowania funkcji f_{obs}).

[35] Ten sposób wyboru usługodawcy jest nierekomendowany przez twórcę systemu RefTRM jako, że może doprowadzić do przeciążenia pewnych agentów w środowisku i nieuzasadnionej dyskryminacji innych agentów.
[36] Jeżeli agent usługodawca nie stosuje ataku.

4.2.9.7. Ocena rekomendacji

Rekomendacje są oceniane po każdej interakcji, co prowadzi do zaktualizowania wartości zaufania rekomendacyjnego (w kontekście c_2) – jeżeli rekomendacja dotycząca usługodawcy była prawidłowa, to zaufanie rekomendacyjne zostaje zwiększone, w przeciwnym przypadku zostaje zmniejszone[37]:

$$f_{trust}(a_i, a_j, c_2, m_l) = t_{a_i \to a_j}^{c_2;m_l} = \begin{cases} \min\left(t_{a_i \to a_j}^{c_2;m_{l-1}} + \alpha^R, 1\right), jeżeli: \left| r_{a_i \to a_j : a_p}^{c_1;m_{l-1}} - o^l \right| \leq h^R \\ \max\left(t_{a_i \to a_j}^{c_2;m_{l-1}} - \beta^R, 0\right), jeżeli: \left| r_{a_i \to a_j : a_p}^{c_1;m_{l-1}} - o^l \right| > h^R \end{cases}$$

gdzie: m_l - moment bezpośrednio po interakcji l; m_{l-1} - moment przed interakcją l (np. po interakcji $l-1$); $\alpha^R \in (0,1)$ – parametr systemu TRM: zwiększenie zaufania rekomendacyjnego za prawidłową rekomendację; $\beta^R \in (0,1)$ – parametr systemu TRM: zmniejszenie zaufania rekomendacyjnego za nieprawidłową rekomendację; $h^R \in (0,1)$ – parametr systemu TRM: tzw. próg prawidłowości rekomendacji, a_p – agent, którego dotyczyła rekomendacja i z którego usług skorzystał usługobiorca.

Czyli jeżeli różnica pomiędzy otrzymaną rekomendacją i rzeczywistym wynikiem interakcji jest nie większa niż próg prawidłowości rekomendacji to zaufanie rekomendacyjne do agenta, który wydał rekomendację, jest zwiększane o α^R, a w przeciwnym przypadku jest zmniejszane o β^R.

4.2.9.8. Ocena lub aktualizacja zaufania lub reputacji

Aktualizacja wartości zaufania w kontekście c_1 (zaufania akcyjnego) następuje po każdej interakcji. Aktualizacji własnego zaufania akcyjnego dokonuje usługobiorca, zgodnie z następującym wzorem:

$$f_{trust}(a_i, a_j, c_1, m_l) = t_{a_i \to a_j}^{c_1;m_l} = \begin{cases} \min\left(t_{a_i \to a_j}^{c_1;m_{l-1}} + \alpha^A, 1\right), jeżeli: o^l \geq h^A \\ \max\left(t_{a_i \to a_j}^{c_1;m_{l-1}} - \beta^A, 0\right), jeżeli: o^l < h^A \end{cases}$$

gdzie: m_l – moment bezpośrednio po interakcji l; m_{l-1} – moment przed interakcją l (moment interakcji $l-1$); $\alpha^A \in (0,1)$ – parametr systemu TRM: zwiększenie zaufania akcyjnego za satysfakcjonującą jakość usługi; $\beta^A \in (0,1)$ – parametr systemu TRM: zmniejszenie zaufania akcyjnego za niesatysfakcjonującą jakość usługi; $h^A \in (0,1)$ – parametr systemu TRM: tzw. próg satysfakcjonującej usługi.

[37] W oryginalnej publikacji [45] warunek, w którym zaufanie jest zmniejszane jest inaczej określony ze względu na niezamierzony błąd. Zaprezentowano wzór zgodnie z intencją autora i pierwszą implementacją systemu RefTRM.

Czyli jeżeli usługa miała satysfakcjonującą jakość (niemniejszą niż określony próg), następuje zwiększenie zaufania o α^A, w przeciwnym przypadku następuje zmniejszenie zaufania o β^A.

4.2.9.9. Podsumowanie

Warto zwrócić uwagę, że sam system RefTRM wprowadził pewne parametry, które mogą wpływać na jego funkcjonowanie, te parametry to: $\alpha, h^A, h^R, h, \alpha^A, \beta^A, \alpha^R, \beta^R$ oraz $t_{init}^{c_1}$ i $t_{init}^{c_2}$, dopiero po określeniu ich wartości w środowisku może funkcjonować system TRM i mogą być przeprowadzane jego badania.

Niniejszy przykład pokazuje, że jest możliwe opisanie systemu RefTRM za pomocą przedstawionego modelu środowiska i modelu systemu TRM, z uwzględnieniem wszystkich aspektów tego konkretnego systemu.

W przypadku innych systemów TRM, przedstawiony model także powinien być wystarczająco pojemny do odzwierciedlenia najistotniejszych właściwości tych systemów.

4.2.10. Ograniczenia modelu

Mimo tego, że przedstawiony model systemu TRM oddaje, zdaniem autora, charakterystykę istotnych cech takich systemów, to, jak każdy model, nie jest w stanie uwzględnić wszystkich aspektów niektórych rzeczywistych systemów. W poniższych punktach opisano pokrótce ograniczenia modelu.

4.2.10.1. Odmowa świadczenia usługi

Zgodnie z przedstawionym modelem, usługodawca nie jest uprawniony do odmowy świadczenia usługi dla konkretnego agenta po tym jak ten ją zażąda. Takie działanie może być uzasadnione w przypadku niektórych typów środowisk, ale samo w sobie oznacza pewien rodzaj dyskryminacji. Można uwzględnić ten aspekt, w ten sposób, że w modelu zostanie dodany krok odpytywania agentów świadczących usługę o to czy są w stanie ją wyświadczyć dla konkretnego agenta. W przedstawionym modelu autor zdecydował się na pominięcie tego aspektu, ze względu na to, że możliwość takiego działania jest rzadka, a z drugiej strony, jak zauważono, sama w sobie jest pewnego rodzaju dyskryminacją i powinna być zwalczana (tak się dzieje w ramach modelu – bo jeżeli agent nie wyświadczy usługi zgodnie ze swoją deklaracją, to usługobiorca obniży zaufanie do niego lub jego reputację).

4.2.10.2. Dostosowanie jakości usługi

Model nie przewiduje, że usługodawca jest uprawniony do dostosowywania jakości usługi do konkretnego usługobiorcy (np. ze względu na jego zaufanie). Takie podejście jest o tyle uzasadnione, że w większości przypadków takie działania nie mają miejsca.

Dodatkowo model nie obejmuje także potencjalnie interesującego zagadnienia dotyczącego możliwego do wyobrażenia procesu negocjacji jakości usługi (a także w wariancie jeszcze szerszym „wynagrodzenia" za jej wyświadczenie) pomiędzy usługodawcą a usługobiorcą.

Możliwe jest jednak rozszerzenie modelu o ten aspekt, aczkolwiek w perspektywie realizacji założeń niniejszej pracy nie wydaje się to konieczne.

4.2.10.3. Możliwość zastosowania w przypadku protokołów routingu

Obecnie model w dość skomplikowany sposób umożliwia odzwierciedlenie usług polegających na przekazywaniu pakietów do innych agentów, w szczególności w przypadku gdy przekazanie danego pakietu wymaga współdziałania wielu agentów. Możliwe jest zastosowanie obecnego modelu w przypadku niektórych protokołów routingu, ale w takim przypadku wzrasta stopień skomplikowania i liczba usług, które należy zdefiniować (np. nie wystarczy zdefiniować usługi polegającej ogólnie na przekazywaniu pakietu, ale konkretnej usługi przekazywania pakietu od konkretnego agenta do innego agenta)., Dodatkowo, do skutecznego przekazania pakietu konieczne jest wyświadczenie wielu usług. Model w założeniu miał być wystarczająco ogólny, aby mógł być stosowany w różnych środowiskach, ale może się okazać, że jego zastosowanie w przypadku niektórych protokołów routingu może być problematyczne.

4.2.10.4. Zależności czasowe

Model nie jest w stanie uwzględnić faktu, że niejednokrotnie ocena jakości usługi, a co za tym idzie ocena zaufania może się odbyć po znacznym czasie od interakcji, gdyż samo świadczenie usługi może trwać długo. Prowadzi to do faktu, że mimo tego, że system TRM jest podatny na ataki wykorzystujące taką specyfikę, to wskutek braku możliwości uwzględnienia jej w modelu nie ma możliwości ich wykrycia. Stanowi to istotne ograniczenie modelu w kontekście celu tej rozprawy, jednak jest ono konieczne z uwagi na znaczący wzrost skomplikowania w przypadku próby odzwierciedlenia takiego aspektu w modelu. Warto także

zauważyć, że model nie jest w stanie zamodelować obsługi kilku żądań trwających w tym samym czasie (jako efekt przyjęcia założenia o zerowym czasie trwania interakcji).

4.2.10.5. Wartość poszczególnych usług

Niektóre systemy TRM przypisują poszczególnym usługom pewną wartość (będącą np. kosztem świadczenia danej usługi lub jej ceną). Niniejszy model nie odzwierciedla takiego aspektu usług, w związku z czym nie jest możliwe uwzględnienie specyfiki pewnych systemów TRM, a także ataków jej dotyczących (takich jak np. świadczenie tanich usług o wysokiej jakości w celu zdobycia wysokiego zaufania lub reputacji i jednoczesne świadczenie drogich usług o niskiej jakości w celu maksymalizacji zysku). Jest to z całą pewnością interesujący aspekt i można rozszerzyć model, tak aby go uwzględnić, ale zostało to pominięte ze względu na to, że głównym celem niniejszej pracy jest analiza działań złośliwych agentów, a nie samolubnych.

4.2.10.6. Koszt funkcjonowania systemu TRM

W przypadku niektórych środowisk, koszty funkcjonowania systemu TRM mogą być istotne (np. w przypadku sieci WSN, w których funkcjonowanie systemu TRM może się wiązać z większym zużyciem energii, czy koniecznością wyposażenia agenta w dodatkowe zasoby). Obecnie model nie uwzględnia tego aspektu, skupiając się na efektywności zapobiegania atakom. Uwzględnienie tego aspektu to niewątpliwie interesujący temat badawczy, ale wymagający badania w odniesieniu do konkretnych środowisk i znacząco wykraczający poza przedmiot zainteresowania niniejszej rozprawy.

4.3. MODEL ATAKU

W podrozdziale przedstawiono model ataku w oparciu o tzw. zachowanie jednostkowe oraz model ataku na bazie współpracy. Obydwa modele opisują atak w odmiennej perspektywie – pierwszy z perspektywy pojedynczego złośliwego agenta, a drugi biorąc pod uwagę ewentualną współpracę pomiędzy agentami.

4.3.1. Model ataku w oparciu o zachowanie jednostkowe

Model ataku w oparciu o zachowanie jednostkowe został zaprezentowany w oparciu o wcześniejsze publikacje [83], [106] autora niniejszej rozprawy.

Definicja 4.3.1.1. *Zachowanie jednostkowe to zachowanie agenta związane z obsługą jednego żądania świadczenia usług lub żądania rekomendacji.*

Rozważmy na co może wpłynąć atakujący żeby zaburzyć działanie środowiska, a także to czego nie jest w stanie zrobić. Warto zauważyć, że atakujący nie może wpłynąć na kształt funkcji innych agentów (np. funkcji wyboru agenta, czy funkcji zaufania), ale może w ogólnym przypadku zmienić informacje wydawane przez siebie (takie jak własne rekomendacje), czy zmienić charakterystyki własnego zachowania w stosunku do innych agentów (np. zmienić jakość świadczonych usług), czy próbować fałszować własną tożsamość.

Rozważając sposób funkcjonowania środowiska (także środowiska z systemem TRM), warto zauważyć, że poszczególne funkcje częściowe takie jak: funkcja interakcji f_{int}, funkcja wyboru usługodawcy f_{sel}, funkcja rekomendacji f_{rec}, itd., są określone przez środowisko oraz ewentualnie system TRM i bazują na charakterystyce agentów, ale niektóre agenty mogą ustalać wartości (wyniki) tych funkcji w sposób niezgodny z charakterystyką środowiska i systemu TRM. Wtedy takie agenty będą przeprowadzać atak przeciwko środowisku lub systemowi TRM. Podsumowując, niewłaściwe działanie (atak elementarny) może dotyczyć[38]:

- sposobu wyboru agenta do interakcji – modyfikacja funkcji wyboru usługodawcy f_{sel}, (niezgodność funkcji zaimplementowanej przez agenta w stosunku do tej narzucanej przez system TRM);
- jakości świadczonych usług – modyfikacja funkcji interakcji f_{int} lub f_{intE} (niezgodność funkcji zaimplementowanej przez agenta w stosunku do tej narzucanej przez środowisko lub system TRM – najczęściej obniżenie jakości świadczonej usługi poniżej maksymalnej jakości usługi, świadczonej przez tego agenta);

[38] Teoretycznie agent może zmodyfikować wyniki prawie wszystkich funkcji częściowych określonych w środowisku i systemie TRM (z wyjątkiem zakłócenia, które jest niezależne od niego), ale istotne są tylko te modyfikacje, które w pewien sposób wpływają na inne agenty (jak np. wartość dostarczonej rekomendacji, czy jakość wyświadczonej usługi), pominięte natomiast są te wyniki funkcji częściowych (np. zaufanie, reputacja), z których korzysta jedynie agent obliczający te wyniki. Wartość obliczonego przez agenta zaufania oczywiście może wpływać na rekomendacje przez niego wydawane, ale wtedy jest to atak na wydawaną rekomendację, co zostało uwzględnione.

- wydawanych rekomendacji – modyfikacja funkcji rekomendacji f_{rec} (niezgodność funkcji zaimplementowanej przez agenta w stosunku do tej narzucanej przez system TRM)[39];
- identyfikacji agenta (świadcząc usługi, czy przekazując rekomendacje agenty działają na podstawie deklarowanej przez siebie tożsamości; w sytuacji gdy istnieje możliwość zmodyfikowania własnej tożsamości agent może przeprowadzić atak na identyfikację, deklarując inną tożsamość lub dokonując jej zwielokrotnienia)[40].

W przypadku gdy nie funkcjonuje system TRM to atak może dotyczyć tylko świadczenia usług i identyfikacji, podobnie w przypadku gdy system TRM nie obsługuje rekomendacji – wtedy oczywiście nie są możliwe manipulacje rekomendacjami.

W literaturze zwykle rozważana jest sytuacja, w której złośliwe agenty są usługodawcami, ale w ogólnym przypadku to założenie nie jest słuszne – złośliwe agenty mogą wszakże nie być usługodawcami, a jedynie dostarczać nierzetelne rekomendacje, np. oczerniać wybranych usługodawców.

W dalszej części rozprawy przyjmujemy, że rzetelny agent nie dokonuje wyboru wyniku wartości powyższych funkcji częściowych, czy własnej tożsamości – stosuje wartości otrzymane jako wyniki funkcji narzuconych przez środowisko lub system TRM oraz prawdziwą tożsamość. Złośliwy agent natomiast za każdym razem podejmuje decyzję o wyniku tych funkcji (i ewentualnie własnej tożsamości, jeżeli pozwala na to środowisko) w sposób arbitralny, niezależny od specyfikacji danego środowiska i systemu. Decyzja w każdym z tych obszarów (funkcji częściowych i tożsamości agenta) może być podejmowana niezależnie.

Definicja 4.3.1.2. *Atakiem elementarnym jest wyznaczenie wyniku przynajmniej jednej z funkcji częściowych w sposób niezgodny z tym określonym w środowisku lub systemie TRM lub niepoprawna identyfikacja agenta.*

[39] Teoretycznie powinniśmy rozważyć także niezgodność funkcji wyboru dostawców rekomendacji f_{sel_rec}, ale z uwagi na to, że z rekomendacji będzie korzystał jedynie agent żądający rekomendacji, (a więc atakujący) aspekt ten został pominięty jako nieistotny.

[40] Stosowane mechanizmy uwierzytelnienia w danym środowisku wpływają na możliwość tego ataku. Sam atak to jednak kwestia deklarowanej tożsamości. W wielu przypadkach, w środowiskach wykorzystujących systemy TRM nie ma mechanizmów uwierzytelnienia, a nawet jeśli są (np. allegro), to i tak istnieje możliwość stworzenia nowego konta (atak na identyfikację) i prawidłowe się uwierzytelnienie (brak ataku na uwierzytelnienie). Takie działanie może być prostsze niż złamanie lub wykradzenie hasła. Atak ten jest wymierzony bezpośrednio w środowisko i nie powinien być rozpatrywany jako atak na system TRM jako taki. Zabezpieczenie przed atakami przeciwko identyfikacji także leży poza możliwościami samego systemu TRM (należy za to do środowiska), ale jest ten problem tu rozważany, ponieważ dotyczy ataków na systemy TRM i w wielu pozycjach literaturowych jest przywoływany w tym kontekście.

Uwaga 4.3.1.1. Przykładem ataku elementarnego może być:

- wydanie nierzetelnej rekomendacji w stosunku do pojedynczego agenta;
- wyświadczenie usługi o jakości niższej niż maksymalna jakość usługi świadczonej przez tego agenta;
- kreacja wielu tożsamości przez agenta.

W środowisku, w którym funkcjonuje system zarządzania zaufaniem i reputacją, agent złośliwy może podejmować następujące rodzaje decyzji (i tym samym wybrać dane zachowanie jednostkowe) w czasie obsługi pojedynczego żądania[41], dla poszczególnych możliwości zachowania wprowadźmy następujące oznaczenia:

- decyzja dotycząca wyboru usługodawcy – wynik funkcji częściowej f_{sel}:
  - zgodna z funkcją określoną dla środowiska lub systemu TRM: ozn. $\boldsymbol{S}^{*}$,
  - niezgodna ozn. $\boldsymbol{S}^{\neq}$, np. ciągłe wybieranie tego samego agenta jako dostawcę usługi[42];
- decyzja dotycząca świadczenia usługi – wynik funkcji częściowej f_{int} lub f_{intE}:
  - świadczenie usługi o zwykłej jakości – ozn. $\boldsymbol{Q}^{*}$ (jakość świadczonej usługi nie musi być obiektywnie dobra, ale jest ona zgodna z określoną przez środowisko lub system TRM funkcją interakcji, np. w większości przypadków jest równa maksymalnej jakości świadczonej usługi przez agenta – agent nie podejmuje prób manipulowania nią),
  - świadczenie usługi o niższej jakości lub brak usługi[43] – ozn. $\boldsymbol{Q}^{-}$;
- decyzja dotycząca wydawania rekomendacji (w przypadku gdy w danym systemie nie występują rekomendacje, nie jest podejmowana taka decyzja) – wynik funkcji częściowej f_{rec}:
  - wydawanie prawidłowych rekomendacji – ozn. $\boldsymbol{R}^{*}$,

[41] Warto podkreślić, że dana decyzja, np. dotycząca wydawania rekomendacji, może mieć konsekwencje w stosunku do wielu żądań. Dzieje się tak w przypadku gdy system TRM nie pozwala na wydawanie rekomendacji przed każdą interakcją.

[42] Może to stanowić pewnego rodzaju atak typu DoS – Denial of Service, poprzez przeciążenie agenta, w szczególności jeżeli środowisko jest na niego podatne.

[43] Brak symetrycznego przypadku dotyczącego zawyżenia jakości usług, ze względu na to że przyjmujemy założenie, że agent nie może sztucznie zawyżać jakości usługi ponad jakość jaką w istocie ma możliwość świadczenia – jest to założenie zgodne z intuicją.

- wydawanie nieprawidłowych rekomendacji – ozn. $\boldsymbol{R^{+}}$ (w przypadku zawyżania rekomendacji), $\boldsymbol{R^{-}}$ (w przypadku zaniżania rekomendacji), $\boldsymbol{R^{\pm}}$ (w przypadku zaniżania lub zawyżania),
- brak rekomendacji – ozn. $\boldsymbol{R^{0}}$ (o ile system TRM dopuszcza możliwość niewydania rekomendacji, a agent byłby w stanie ją wydać, ale ze względu na własne strategiczne działanie tego nie robi);

- decyzja dotycząca identyfikacji:
  - identyfikacja prawidłowa (swoim identyfikatorem), lub w danym środowisku w ogóle nie jest dokonywana identyfikacja agentów) – ozn. $\boldsymbol{I^{*}}$,
  - identyfikacja nieprawidłowa (nieswoim identyfikatorem – stosowanie tzw. spoofingu) – ozn. $\boldsymbol{I^{\neq}}$,
  - kreacja wielu tożsamości - ozn. $\boldsymbol{I^{\times}}$**.**

Oczywiście, nie przy każdym żądaniu agent złośliwy podejmuje wszystkie decyzje, w szczególności gdy podejmuje decyzję o wyborze usługodawcy to nie podejmuje decyzji dotyczącej jakości świadczenia usługi. Co więcej, konieczność podjęcia decyzji dotyczącej identyfikacji być może zaistniała wcześniej (na początku działania środowiska) i nie ma możliwości jej zmiany w pojedynczym żądaniu – zależy to od specyfiki środowiska. Ponadto, w systemie TRM mogą nie występować rekomendacje (albo w danym środowisku może nie być systemu TRM), lub rekomendacje mogą być wydawane nie w odniesieniu do konkretnego żądania, więc decyzja dotycząca rekomendacji także może nie być podejmowana.

Należy zauważyć, że nie w każdym aspekcie działanie musi być określone, ponieważ dany atak może nie specyfikować zachowania agenta w określonym aspekcie, co więcej w przypadku niektórych środowisk pewne zachowania agenta mogą nie być w ogóle możliwe. Z tego względu wprowadźmy następujące oznaczenia:

- w przypadku gdy dane działanie nie jest charakterystyczne dla ataku, albo w ramach różnych wariantów ataku, agent może postępować w różny sposób w zależności od sytuacji, przyjmujemy oznaczenie: ?
- w przypadku gdy agent w ogóle nie podejmuje takiej decyzji (np. ze względu na specyfikę swojej charakterystyki – np. jest tylko usługodawcą, lub charakterystykę systemu TRM – np. nie są stosowane rekomendacje), przyjmujemy oznaczenie: –

Warto zwrócić uwagę, że nieco podobny sposób rozumowania w celu analizy możliwych zachowań złośliwych agentów (jednak mniej rozbudowany) został przedstawiony w artykule [89].

W czasie obsługi jednego żądania[44] agent może mieć okazję podjąć opisane powyżej cztery decyzje (nie wszystkie jednocześnie), ale co najwyżej jedną decyzję spośród dwóch pierwszych decyzji (czyli wybór usługodawcy i jakość świadczonej usługi), w zależności od tego czy w danej interakcji jest usługodawcą czy usługobiorcą[45]. Wprowadźmy następujące oznaczenia:

- $(.../\ ...,\ ...,\ ...)$ – oznacza charakterystykę zachowania w związku z obsługą jednego żądania, np.:
  - $(\boldsymbol{S}^*/\boldsymbol{Q}^*, \boldsymbol{R}^*, \boldsymbol{I}^*)$ – oznacza, że w związku z obsługą danego żądania agent postąpił rzetelnie (wykonał rzetelne działanie w odniesieniu do każdej decyzji);
- $[.../\ ...,\ ...,\ ...]$ – oznacza charakterystykę zachowania pewnego agenta, np.:
  - $[\boldsymbol{S}^*/\boldsymbol{Q}^*, \boldsymbol{R}^*, \boldsymbol{I}^*]$ – oznacza, że agent zawsze postępuje rzetelnie,
  - $[\boldsymbol{S}^*/\boldsymbol{Q}^*, \boldsymbol{R}^{\pm}, \boldsymbol{I}^*]$ – oznacza, że agent zawsze, kiedy ma taką możliwość, udziela nieprawidłowych rekomendacji,
  - $[\boldsymbol{S}^*/\boldsymbol{Q}^*, \dddot{\boldsymbol{R}}^{\pm}, \boldsymbol{I}^*]$ – oznacza, że agent może wydawać nieprawidłowe rekomendacje[46].

Stosując powyższe oznaczenia można dokonać opisu każdego ataku w ramach modelu w oparciu o zachowania jednostkowe agentów. Dodatkowo należy zauważyć, że niektóre z ataków opisanych w literaturze mogą być stosowane łącznie, tworzą odmienną charakterystykę ataku i odmienny opis w ramach tego modelu. Stworzenie takiego modelu jest istotne gdyż w wielu przypadkach w literaturze, mimo stosowania podobnej, czy tej samej nazwy ataku, zakładane jest inne zachowanie agentów.

[44] W mocy pozostaje stwierdzenie dokonane wcześniej. że ze względu na sposób działania systemu TRM, agent może nie być w stanie podejmować danej decyzji selektywnie w trakcie każdej interakcji, a może być zobligowany wcześniejszymi swoimi decyzjami.

[45] W ramach danego żądania agent może nie mieć możliwości podjęcia żadnej z tych dwóch decyzji, ponieważ interakcja odbywa się pomiędzy innymi agentami, ale w takim przypadku być może wciąż będzie miał możliwość wydania rekomendacji, czyli podjęcia decyzji dotyczącej rekomendacji, choć i tej decyzji może być także pozbawiony.

[46] Oznaczenie, że w danym ataku atakujący stosują zachowanie $\dddot{\boldsymbol{R}}^{\pm}$ nie oznacza, że atakujący zawsze wydają nieprawidłowe rekomendacje, a jedynie, że robią to w pewnych przypadkach (np. w stosunku do określonych agentów, lub co określony czas).

4.3.1.1. Specyfikacja ataku

Do wyspecyfikowania konkretnego ataku konieczne jest określenie zachowania agenta w odniesieniu do każdego z czterech podanych powyżej aspektów, stosując powyższe oznaczenia, poszczególne symbole oznaczeń powinny być wymienione w następującej kolejności:

[zachowanie dotyczące wyboru usługodawcy/zachowanie dotyczące świadczenia usług, zachowanie dotyczące rekomendacji, zachowanie dotyczące identyfikacji]

Istotne w tym modelu jest to, że agent dokonuje modyfikacji określonych funkcji częściowych, określając je w sposób odmienny od sposobu założonego w środowisku, ale model nie uwzględnia dokładnej charakterystyki w jakich przypadkach określone zachowanie jednostkowe jest stosowane, w przypadku gdy agent nie zawsze stosuje dane zachowanie jednostkowe. Co więcej, model też nie specyfikuje jakie wartości zmodyfikowanych funkcji (np. funkcji rekomendacji) stosuje agent w czasie ataku. Dlatego ta charakterystyka powinna być określona dodatkowo.

4.3.1.2. Przykłady opisu ataków

Dzięki uwzględnieniu zachowania jednostkowego agenta w kontekście podejmowanych czterech decyzji, istnieje możliwość opisania każdego ze zidentyfikowanych w literaturze ataków (opis poszczególnych ataków został dokonany w załączniku 4), np.:

- atak wychwalania lub oczerniania: $[\boldsymbol{S}^{*}/\boldsymbol{Q}^{*}, \boldsymbol{R}^{\pm}, \boldsymbol{I}^{*}]$, przy czym:
  - atak wychwalania: $[\boldsymbol{S}^{*}/\boldsymbol{Q}^{*}, \boldsymbol{R}^{+}, \boldsymbol{I}^{*}]$,
  - atak oczerniania: $[\boldsymbol{S}^{*}/\boldsymbol{Q}^{*}, \boldsymbol{R}^{-}, \boldsymbol{I}^{*}]$,
- atak kreacji wielu tożsamości $[?/?, ?, \boldsymbol{I}^{\times}]$,
- atak wybielania $[?/\ddot{\boldsymbol{Q}}^{\ddot{-}}, ?, \boldsymbol{I}^{\neq}]$
- atak stały $[?/\boldsymbol{Q}^{-}, ?, \boldsymbol{I}^{*}]$,
  - atak stały w systemie TRM, który nie wykorzystuje rekomendacji: $[?/\boldsymbol{Q}^{-}, -, \boldsymbol{I}^{*}]$,
- atak oscylacji zachowania (on-off): $[?/\ddot{\boldsymbol{Q}}^{\ddot{-}}, ?, \boldsymbol{I}^{*}]$, przy czym:
  - atak oscylacji zachowania w systemie TRM, który nie wykorzystuje rekomendacji: $[?/\ddot{\boldsymbol{Q}}^{\ddot{-}}, -, \boldsymbol{I}^{*}]$.

4.3.1.3.Ograniczenia modelu

W poniższych punktach opisano pokrótce ograniczenia modelu.

4.3.1.3.1. Brak uwzględnienia współpracy wielu agentów

Model zakłada, że dokonywany jest opis zachowania jednego agenta, a więc nie ma możliwości opisania współpracy pomiędzy agentami inaczej niż poprzez niezależne zachowania tych agentów. W takim przypadku jednak mogą zostać pominięte istotne aspekty tej współpracy, szczególnie w przypadku gdy atakujący działają w sposób skoordynowany i wysublimowany. Dlatego też współpraca pomiędzy wieloma agentami musi zostać opisana w odmienny sposób.

4.3.1.3.2. Brak uwzględnienia specyficznych aspektów ataku

Częścią charakterystyki ataku może być zmienność jego przebiegu w czasie, np. wykonywanie przez agenta nierzetelnego działania z pewnym prawdopodobieństwem, albo stosowanie niewłaściwej rekomendacji jedynie w stosunku do wybranych agentów. Model sam w sobie nie określa też dokładnego sposobu modyfikacji (wartości) poszczególnych funkcji częściowych. Aspekty te muszą być określone dodatkowo przy wyspecyfikowaniu ataku.

4.3.2. Model ataku na bazie współpracy

W niniejszym punkcie zaprezentowano propozycję innego modelu ataku – na bazie współpracy. Potrzeba stworzenia drugiego modelu wynika z faktu, że model ataku na w oparciu o zachowanie jednostkowe jest bardzo prosty, nie uwzględnia między innymi możliwości współpracy agentów przy wykorzystaniu ataku. Z tego względu rozważmy następujące założenia – analizę sytuacji, jaka może zachodzić w środowisku zarówno z perspektywy rzetelnego agenta, jak i złośliwego agenta.

Założenia:

- w środowisku występuje pewna liczba rzetelnych agentów i złośliwych agentów, przy czym może wystąpić sytuacja, że liczba rzetelnych agentów, lub liczba złośliwych agentów jest równa 0,
- rzetelny agent nie wie czy w systemie są złośliwe agenty,
- rzetelny agent nie wie ile jest złośliwych agentów,
- rzetelny agent nie wie nic na temat grup złośliwych agentów(nie wie ile ich jest, jaka jest ich liczebność, itd.),
- rzetelny agent nie wie czy określony inny agent jest złośliwym agentem czy rzetelnym agentem,

- złośliwy agent wie, które agenty należą do jego grupy,
- złośliwy agent nie wie czy istnieją inne grupy złośliwych agentów, ile jest takich grup oraz jaka jest ich liczebność i które agenty do nich należą, w szczególności nie wie, czy określony inny agent spoza jego grupy jest agentem rzetelnym, czy złośliwym.

Z powyższych założeń wynika, że w takim środowisku występuje asymetria informacji – agenty złośliwe dysponują większą wiedzą niż agenty rzetelne (wiedzą, ile i jakie agenty należą do tej samej złośliwej grupy agentów). W przypadku gdy występuje więcej grup nierzetelnych agentów (niewspółpracujących ze sobą), to wielkość asymetrii informacji się zmniejsza. Rozważenie wielu niezależnych grup złośliwych agentów wprowadza większą niepewność co do zachowania innych agentów, przy czym paradoksalnie zwiększenie niepewności jest większe dla agentów złośliwych. Wynika to z faktu, że w przypadku jednej grupy złośliwych agentów, dany złośliwy agent zawsze wiedział jak zachowa się jego partner interakcji (z dokładnością do zakłóceń, występujących w środowisku). W momencie gdy pojawia się druga grupa złośliwych agentów, żaden ze złośliwych agentów nie wie już jak się zachowa agent, z którym podejmowana jest interakcja. Z drugiej strony, rzetelne agenty tracą niewiele (wzrasta całkowita liczba złośliwych agentów). Wtedy jest możliwość nawiązania interakcji pomiędzy parą złośliwych agentów należących do różnych grup, co może doprowadzić do załamania strategii którejś z grup atakujących (być może wszystkich).

W literaturze są zwykle rozważane dwa przypadki:

- w środowisku istnieje wiele pojedynczych, niezależnych złośliwych agentów,
- w środowisku istnieje jedna grupa współpracujących ze sobą złośliwych agentów.

Wedle najlepszej wiedzy autora, nie istnieje praca, która rozważałaby istnienie w środowisku kilku niezależnych grup współpracujących ze sobą (w ramach grupy) agentów i oceniała wpływ takiej sytuacji na efektywność działań atakujących, a tego typu zagadnienie wydaje się być interesującym tematem badawczym.

Analizując współpracę agentów w ramach grupy, można stwierdzić, że działania agentów mogą być:

- wspólne (agenty wykonują zawsze ten sam atak elementarny w odniesieniu do pojedynczego żądania – np. dostarczają zaniżoną rekomendację na temat innych agentów);
- zsynchronizowane (agenty wykonują różne działania w odniesieniu do jednego żądania usługobiorcy, np. jeden złośliwy agent dostarcza zawyżoną rekomendację, a drugi

świadczy usługę o złej jakości, ale są one wzajemnie ustalone i skoordynowane przez złośliwe agenty w grupie);

- niezależne[47] (agenty wykonują niezależne działania, które nie są wzajemnie ustalone i skoordynowane przez złośliwe agenty w grupie).

W modelu przyjęto założenie o całkowitej niezależności grup złośliwych agentów, ale warto wskazać, że poszczególne grupy agentów mogą także być:

- współpracujące (grupy znają nawzajem swoją całą charakterystykę, obejmująca liczbę agentów, wykonywany atak itd. i wspólnie ustalają bieżące działania);
- skoordynowane (grupy agentów nie znają całej charakterystyki innej grupy, np. nie wiedzą które węzły przynależą do grupy, ale wspólnie ustalają działania);
- świadome (grupy nie ustalają wspólnych działań, ale znają część swojej charakterystyki, np. przynależność do grupy);
- niezależne.

4.3.2.1. Specyfikacja ataku

Opisując atak, konieczne jest określenie, ile grup złośliwych agentów jest zaangażowanych w atak, w jaki sposób agenty współpracują ze sobą w ramach poszczególnych grup oraz w jaki sposób grupy współpracują ze sobą (jeżeli grup złośliwych agentów jest więcej niż jedna).

4.3.2.2. Przykład opisu ataków

W przypadku ataku wyrocznia (opisanego w załączniku 4), występują dwie grupy złośliwych agentów: obserwatorzy i wyrocznie. W ramach każdej grupy działania są wspólne, a grupy działają w sposób co najmniej skoordynowany.[48]

4.3.2.3. Ograniczenie modelu – brak uwzględnienia specyficznych aspektów ataku

Podobnie jak model ataku w oparciu o zachowanie jednostkowe, tak i ten model nie odzwierciedla wszystkich specyficznych działań atakujących, ani nie jest w stanie ich jednoznacznie opisać. Wszakże model ten stanowi jakoby klasyfikację możliwych form współpracy pomiędzy wieloma agentami.

---

[47] Z punktu widzenia modelu nie ma znaczenia czy poszczególne agenty potraktujemy jako należące do oddzielnych grup – zawierających jednego agenta, czy też grupę agentów działających w pełni niezależnie. Wydaje się że jest sens traktować agenty jako oddzielne grupy w sytuacji gdy stosują odmienne zachowanie w kontekście ataku elementarnego

[48] Ten sam atak można także opisać inaczej: można przyjąć, że zarówno obserwatorzy, jak i wyrocznie należą do jednej grupy. Wtedy w ramach grupy, agenty wykonują zadania zsynchronizowane.

5. METODYKA OCENY WIARYGODNOŚCI SYSTEMÓW TRM

W niniejszym rozdziale zaprezentowano metodykę oceny wiarygodności systemów zarządzania zaufaniem i reputacją. W poszczególnych podrozdziałach zaprezentowano, miary wiarygodności, rodzaje oraz propozycje badań, które zdaniem autora, powinny być wykorzystywane do oceny wiarygodności systemów TRM. Głównym celem niniejszej pracy nie jest całościowa ocena systemów TRM, ale ocena ich odporności na ataki. Oznacza to, że nie będą szczegółowo brane pod uwagę aspekty związane z kosztem funkcjonowania samego systemu TRM (mierzone np. koniecznością przeprowadzania dodatkowych obliczeń czy komunikacji).

5.1. MIARY WIARYGODNOŚCI

Poniższe miary wiarygodności mają zastosowanie do wszystkich systemów zarządzania zaufaniem i reputacją. Zaproponowane miary w niektórych przypadkach (np. efektywność) pokrywają się w pewnym stopniu z tymi zdefiniowanymi w literaturze [69], [90], [91]. Część wymienionych miar, została wcześniej określona, w nieco innym ujęciu, w innych publikacjach autora rozprawy [66], [79], [107].

5.1.1. Miary efektywności

Miary efektywności środowiska określają jaka jest sprawność działania środowiska i na ile jest ono odporne na działanie złośliwych agentów.

Definicja 5.1.1.1. ***Efektywność środowiska bez zakłóceń (E_q)*** *– stosunek sumy wyników interakcji do iloczynu liczby wszystkich interakcji i maksymalnej jakości usług w środowisku:*

$$E_q = \frac{\sum_{i=1}^{l} q^i}{l * q_{max}}$$

Definicja 5.1.1.2. ***Efektywność środowiska (E)*** *– stosunek sumy rzeczywistych wyników interakcji do iloczynu liczby wszystkich interakcji[49] i maksymalnej jakości usług w środowisku:*

$$E = \frac{\sum_{i=1}^{l} o^i}{l * q_{max}}$$

W powyższych wzorach l jest liczbą interakcji w środowisku.

[49] W przypadku gdy $O = \{0,1\}$, efektywność środowiska to liczba interakcji zakończonych sukcesem przez liczbę wszystkich interakcji.

Definicja 5.1.1.3. ***Efektywność chwilowa n ($E^{(n)}$)*** *– jest to efektywność środowiska, ale biorąca pod uwagę jedynie n ostatnich interakcji*[50]*:*

$$E^{(n)} = \frac{\sum_{i=l-n+1}^{l} o^i}{n * q_{max}}$$

W powyższym wzorze l jest numerem interakcji, dla której jest wyznaczana miara efektywności chwilowej n.

Uwaga 5.1.1.1. W szczególności, **efektywność chwilowa** to efektywność chwilowa 1 ($E^{(1)}$):

$$E^{(1)} = \frac{o^l}{q_{max}}$$

W powyższym wzorze l jest numerem interakcji, dla której jest wyznaczana efektywność chwilowa.

Miara analogiczna do efektywności środowiska jest często stosowana w literaturze, z dokładnością do braku wartości q_{max} w mianowniku wzoru, co wynika z faktu, że maksymalna jakość usług przyjmowana zwykle w modelach ma wartość 1, przy czym rzadko kiedy jest to określone explicite. Miara ta, choć powszechnie stosowana i stanowiąca obraz funkcjonowania środowiska, ma poważną wadę w odniesieniu do oceny wpływu ataków na to środowisko. Warto bowiem zwrócić uwagę na fakt, że w rzeczywistym środowisku może nie istnieć agent, który daną usługę (np. usługę u_l) będzie świadczyć z maksymalną jakością usług, tzn. $\nexists k: q_{a_k}^{u_l} = q_{max}$. Wobec tego, nawet przy braku jakichkolwiek złośliwych działań ze strony atakujących, efektywność środowiska bez zakłóceń (a w sytuacji gdy w środowisku nie występują zakłócenia także efektywność środowiska), będzie mniejsza od jedności. Uzasadnionym jest więc wprowadzenie kolejnych dwóch miar.

Zauważmy, że dla każdej usługi w środowisku może być wyznaczona maksymalna jakość usługi z jaką może być świadczona, tzn. dla usługi u_l: $q_{max}^{u_l} = q_{a_k}^{u_l}: \forall a_j \in A_{P:u_l}: q_{a_j}^{u_l} \leq q_{a_k}^{u_l}$. Biorąc to pod uwagę zdefiniujmy kolejną miarę: idealną efektywność.

Idealna efektywność to efektywność środowiska, w którym usługi byłyby świadczone jedynie przez agentów, którzy dostarczają usługi o najlepszej jakości w środowisku i w sytuacji, w której żaden z nich nie wykonuje żadnego ataku. Innymi słowy, efektywność taka zakłada,

[50] Warto zauważyć, że analogiczną miarę do efektywności chwilowej można stworzyć pod kątem efektywności środowiska bez zakłóceń, poprzez zastąpienie rzeczywistego wyniku interakcji o_l, wynikiem interakcji q_l.

że usługobiorcy wybiorą zawsze najlepszego dostawcę usług. Efektywność idealna jest równa 1 w sytuacji gdy agenci nie wykonują żadnego ataku w odniesieniu do świadczenia usług, a usługobiorcy idealnie wybierają usługodawcę. Przyjmijmy następującą konwencję oznaczenia: $u_{int:l}$ to usługa, która była świadczona w ramach l – tej interakcji.

Definicja 5.1.1.4. ***Idealna efektywność (E^{ideal})*** *jest równa stosunkowi sumy rzeczywistych wyników interakcji do sumy maksymalnej jakości usług w całym środowisku świadczonych przez agenty w kolejnych interakcjach.*

$$E^{ideal} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{max}^{u_{int:i}}}$$

Przy czym $q_{max}^{u_{int:i}}$ to maksymalna jakość usługi (w całym środowisku), która była świadczona podczas i – tej interakcji.

Idealna efektywność bierze pod uwagę globalną (dla całego środowiska) jakość usług, ale przy obliczaniu miary efektywności możemy też wziąć pod uwagę maksymalną jakość usługi agenta, który świadczył tę usługę podczas określonej interakcji. Zaawansowana efektywność bierze pod uwagę, które agenty świadczyły daną usługę. Przyjmijmy następującą konwencję: $a_{int:l}$ to agent, który świadczył usługę w ramach l – tej interakcji, a $u_{int:l}$, to usługa, która była świadczona podczas tej interakcji.

Definicja 5.1.1.5. ***Zaawansowana efektywność (E^{adv})*** *jest równa stosunkowi sumy rzeczywistych wyników interakcji do sumy maksymalnej jakości usług danego agenta świadczonych w kolejnych interakcjach:*

$$E^{adv} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{a_{int:i}}^{u_{int:i}}}$$

Przy czym $q_{a_{int:i}}^{u_{int:i}}$ to maksymalna jakość usługi świadczonej przez agenta, który był usługodawcą podczas i – tej interakcji.

Analizując wzory miar efektywności środowiska, idealnej efektywności oraz zaawansowanej efektywności, można zauważyć następującą zależność:

Lemat 5.1.1. Zawsze zachodzi następująca relacja pomiędzy różnymi typami efektywności:

$$E^{adv} \geq E^{ideal} \geq E$$

Dowód: z definicji zachodzi: $q_{a_{int:i}}^{u_{int:i}} \leq q_{max}^{u_{int:i}} \leq q_{max}$, dla każdego i, co implikuje $E^{adv} \geq E^{ideal} \geq E$ (ze względu na równość licznika we wzorach na poszczególne rodzaje efektywności).

Zauważmy, że efektywność środowiska może wynosić 1 tylko w sytuacji gdy dla każdej usługi żądanej, istnieje przynajmniej jeden usługodawca, który jest w stanie świadczyć tę usługę z maksymalną jakością usług w systemie.

Przykład 5.1.1.1: Rozważmy środowisko składające się z trzech agentów usługodawców $A_P = \{a_1, a_2, a_3\}$, przy czym każdy z tych agentów świadczy usługę u_1. W środowisku zdefiniowano następujący zbiór możliwych jakości usług: $Q = \{0;\ 0{,}25;\ 0{,}5;\ 0{,}75;\ 1\}$, przy czym każdy z agentów może świadczyć usługę u_1 z określoną maksymalną jakością: $q_{a_1}^{u_1} = 0{,}25,\ q_{a_2}^{u_1} = 0{,}5,\ q_{a_3}^{u_1} = 0{,}75$. Z każdym z tych agentów inne agenty, nawiązały po dwie interakcje, czyli było w sumie 6 żądań świadczenia usług i interakcji: $l = 6$. W środowisku nie występują zakłócenia, czyli $\forall i \in \mathbb{N}, 1 \leq i \leq 6$: $o_i = q_i$. Przyjmijmy, że żaden z agentów usługodawców nie wykonał ataku dotyczącego świadczenia usług, czyli świadczył usługi z maksymalną możliwą jakością określoną przez swoją charakterystykę. Wtedy poszczególne miary efektywności będą wynosić:

$$E_q = E = \frac{\sum_{i=1}^{l} o^i}{l * q_{max}} = \frac{2 * 0{,}25 + 2 * 0{,}5 + 2 * 0{,}75}{6 * 1} = \frac{1}{2}$$

$$E^{ideal} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{max}^{u_{int:i}}} = \frac{2 * 0{,}25 + 2 * 0{,}5 + 2 * 0{,}75}{6 * 0{,}75} = \frac{2}{3}$$

$$E^{adv} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{a_{int:i}}^{u_{int:i}}} = \frac{2 * 0{,}25 + 2 * 0{,}5 + 2 * 0{,}75}{2 * 0{,}25 + 2 * 0{,}5 + 2 * 0{,}75} = 1$$

Przykład 5.1.1.2: Rozważmy środowisko i żądania takie same jak w przykładzie 5.2.1.1., ale w tym przypadku załóżmy, że każdy z agentów przeprowadzi atak dotyczący świadczenia usługi podczas jednej z dwóch własnych interakcji, świadcząc wtedy usługę z jakością $q_i = 0$, dla $i = 2; 4; 6$.[51] Wtedy poszczególne miary efektywności będą wynosić:

$$E_q = E = \frac{\sum_{i=1}^{l} o^i}{l * q_{max}} = \frac{0{,}25 + 0 + 0{,}5 + 0 + 0{,}75 + 0}{6 * 1} = \frac{1}{4}$$

[51] Czyli każdy z tych agentów wykonuje atak oscylacji zachowania (on-off): $[?/\ddot{\boldsymbol{Q}}^{-}, ?, \boldsymbol{I}^{*}]$

$$E^{ideal} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{max}^{u_{int:i}}} = \frac{0{,}25 + 0 + 0{,}5 + 0 + 0{,}75 + 0}{6 * 0{,}75} = \frac{1}{3}$$

$$E^{adv} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{a_{int:i}}^{u_{int:i}}} = \frac{0{,}25 + 0 + 0{,}5 + 0 + 0{,}75 + 0}{2 * 0{,}25 + 2 * 0{,}5 + 2 * 0{,}75} = \frac{1}{2}$$

Najistotniejszym parametrem w kontekście oceny systemów TRM jest idealna efektywność. Wynika to z faktu, że nawet wysoka wartość zaawansowanej efektywności nie musi świadczyć o optymalnych wyborach agentów (wybieraniu tego agenta, który świadczy najlepsze usługi), co jest wszakże głównym celem systemów TRM. Z drugiej strony, używanie miary efektywności środowiska może prowadzić do sytuacji, że mimo niskiej wartości, tak naprawdę decyzje podejmowane przez usługobiorców są optymalne (wynika to z faktu, że jeżeli żaden agent nie świadczy usługi o maksymalnej jakości, to i tak wartość efektywności będzie niższa od jedności). Warto jednak zauważyć, że o ile wyznaczenie efektywności środowiska jest łatwe (wystarczy znać wyniki interakcji), o tyle wyznaczenie pozostałych miar efektywności wiąże się z większymi problemami, wymaga bowiem wiedzy o charakterystyce wszystkich usługodawców uczestniczących w interakcjach (w przypadku zaawansowanej efektywności), czy wręcz charakterystyki wszystkich agentów w systemie i wszystkich żądań (w przypadku efektywności idealnej).

Każda ze zdefiniowanych miar może być wyznaczona nie tylko w odniesieniu do wszystkich żądań (i interakcji), które zaszły w środowisku, ale także może zostać obliczona w każdym momencie działania środowiska, w odniesieniu do interakcji, które zaszły od początku działania środowiska do rozważanego momentu.

5.1.2. Miary zysku efektywności

W bieżącym punkcie zostaną wprowadzone dodatkowe miary dotyczące zysku z istnienia systemu TRM funkcjonującego w środowisku. Do tego celu wprowadźmy pojęcie najbardziej efektywnego ataku.

Definicja 5.1.2.1. ***Najbardziej efektywny atak*** *to taki sposób działania atakujących (podejmowania decyzji w odniesieniu do aspektów wymienionych w modelu ataku), który pozwala zminimalizować efektywność środowiska.*

Koncepcja najbardziej efektywnego ataku jest teoretyczna, a jego wyznaczenie, w szczególności przy nieznajomości a priori charakterystyki wszystkich żądań w systemie, jest wysoce nietrywialne lub niemożliwe[52]. Ponadto, wpływ na to jaki atak jest najbardziej efektywny mają takie czynniki jak funkcja wyboru usługodawcy, czy zastosowany system zarządzania zaufaniem i reputacją (włącznie z wszystkimi wartościami jego parametrów) oraz wszystkie żądania świadczenia usług, włącznie z ich kolejnością. Rozważając środowisko i żądania świadczenia usług, warto zauważyć, że wprowadzenie do takiego środowiska systemu TRM lub zmiana jego parametrów, może spowodować, ceteris paribus, zmianę najbardziej efektywnego ataku. Z tego względu, porównując środowisko bez oraz z systemem TRM, będziemy mieli do czynienia z innym najbardziej efektywnym atakiem, czyli z inną charakterystyką działań atakujących.

Istotnym zagadnieniem jest rozstrzygnięcie w jaki sposób stosowanie systemu TRM (konkretnego systemu TRM, z ustalonymi parametrami) wpływa na efektywność, przy obecności złośliwych agentów przeprowadzających atak. Do oceny takiego wpływu miary nazwane zyskiem efektywności i zyskiem absolutnym efektywności, mogą znaleźć zastosowanie, które zostały zdefiniowane w publikacji [79] autora rozprawy:

Definicja 5.1.2.2. ***Zysk efektywności (G)*** *– różnica pomiędzy efektywnością środowiska, w którym działa określony system zarządzania zaufaniem (E_{+TRM}), a efektywnością środowiska bez systemu zarządzania zaufaniem (E_0), przy założeniu, że w obydwu przypadkach atakujący zachowują się dokładnie w ten sam sposób – stosują działania pozwalające maksymalnie obniżyć efektywność środowiska w momencie korzystania z systemu zarządzania zaufaniem:*

$$G = E_{+TRM} - E_0$$

Definicja 5.1.2.3. ***Zysk absolutny efektywności (G_A)*** *– różnica pomiędzy efektywnością środowiska, w którym działa określony system zarządzania zaufaniem (E_{+TRM}) i efektywnością środowiska bez systemu zarządzania zaufaniem ($E_{0'}$) przy założeniu, że w obydwu przypadkach atakujący stosują najbardziej efektywny atak, tj. tak dobierają swoje działania aby maksymalnie obniżyć efektywność środowiska zarówno w przypadku kiedy system zarządzania zaufaniem jest używany, jak i wtedy gdy nie jest (mogą stosować różne strategie działania w tych dwóch przypadkach):*

$$G_A = E_{+TRM} - E_{0'}$$

[52] Zostanie to przedyskutowane w dalszej części rozdziału.

Uwaga 5.1.2.1 Warto zauważyć, że z definicji wynika, że: $G_A \geq G$, ponieważ: $E^{0\prime} \leq E^0$, ze względu na to, że $E^{0\prime}$ jest efektywnością zmierzoną w sytuacji gdy atakujący stosują najbardziej efektywny atak dostosowany do wszystkich warunków środowiska (w tym do braku istnienia systemu TRM).

Twierdzenie 5.1.2.1. W przypadku gdy w środowisku nie funkcjonuje system TRM, a wybór usługodawców odbywa się w sposób losowy[53], najbardziej efektywnym atakiem jest atak stały połączony z kreacją wielu tożsamości, w którym agenty złośliwe dostarczają usługi o minimalnej możliwej jakości. Wtedy efektywność środowiska (przy założeniu, że wszystkie rzetelne agenty są w stanie świadczyć usługi z jakością q_{max}) wynosi w przybliżeniu[54]:

$$E_{0\prime} \approx \frac{n_B + n_M \frac{q_{min}}{q_{max}}}{n_B + n_M}$$

Gdzie n_M to liczba złośliwych agentów z perspektywy środowiska (po zastosowaniu ataku kreacji wielu tożsamości, czyli uwzględniająca liczbę tożsamości, którymi mogą się identyfikować złośliwe węzły).

Uwaga 5.1.2.2. W sytuacji gdy $q_{min} = 0$, a $q_{max} = 1$ (najbardziej intuicyjny przypadek), efektywność wynosi: $E_{0\prime} \approx \frac{n_B}{n_B + n_M}$

Dowód: Zauważmy, że w takim środowisku jakość dostarczanych usług nie ma wpływu na wybór danego agenta jako usługodawcy. Ponieważ rzeczywisty wynik każdej interakcji wchodzi do licznika wzoru na efektywność, to za każdym razem złośliwy agent jest zainteresowany dostarczeniem usługi o minimalnej jakości. Z drugiej strony, agenty złośliwe mogą zwiększyć częstotliwość wyboru jako usługodawca stosując atak kreacji wielu tożsamości. Wtedy charakter przedstawionych wzorów jest oczywisty.

5.1.3. Miary globalnego średniego zaufania i reputacji

Do oceny odporności systemu TRM na ataki można wykorzystać miary oparte na średnim zaufaniu pomiędzy agentami. W szczególności im wyższe wartości zaufania agentów rzetelnych do innych agentów rzetelnych i im niższe wartości zaufania agentów rzetelnych do złośliwych, tym środowisko powinno funkcjonować lepiej. Miary te mają jednak tę istotną

[53] Częsty sposób funkcjonowania środowisk, w przypadku niewykorzystywania systemu TRM, oczywiście przy założeniu pełnej homogeniczności świadczonych usług (a dokładniej deklaracji dotyczących jakości świadczonych usług).

[54] Przybliżenie wynika z faktu losowości wyboru agentów, wynik jest tym bardziej zbieżny do wskazanej wartości, im więcej interakcji zaszło w środowisku.

wadę, że są trudne do wyznaczenia w rzeczywistym środowisku, w którym problem sprawia jednoznaczne rozstrzygnięcie, które z agentów są nierzetelne. W przypadku symulacji, czy też w środowiskach testowych badających efektywność systemów TRM, miary te są w stanie umożliwić ocenę odporności systemów TRM na ataki.

Definicja 5.1.3.1. ***Globalne średnie zaufanie w kontekście c_k*** *to suma wartości zaufania w kontekście c_k pomiędzy każdą uporządkowaną parą agentów, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania wszystkich agentów do wszystkich agentów $(A \to A)$ w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$t_{A\to A}^{c_k;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n}\sum_{j=1}^{n} 1}$$

przy czym i, j są takie że zaufanie agenta a_i do a_j jest określone.

Uwaga 5.1.3.1. Jeżeli dla każdej pary agentów jest określona wartość zaufania, to globalne średnie zaufanie wszystkich agentów do wszystkich agentów $(A \to A)$ w kontekście c_k w chwili m_l jest wyrażone wzorem:

$$t_{A\to A}^{c_k;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} t_{a_i\to a_j}^{c_k;m_l}}{n(n-1) * t_{max}}$$

Przykład 5.1.3.1. Na podstawie definicji 5.1.3.1. i w odniesieniu do przykładu systemu RefTRM, opisanego w punkcie 4.2.9, warto zauważyć, że w systemie RefTRM można określić następujące miary globalnego średniego zaufania:

- globalne średnie zaufanie w kontekście c_1 (które analogiczne jak zaufanie w tym kontekście można nazwać **globalnym średnim zaufaniem akcyjnym**), równe:

$$t_{A\to A}^{c_1;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} t_{a_i\to a_j}^{c_1;m_l}}{n(n-1)}$$

- globalne średnie zaufanie w kontekście c_2 (które analogiczne jak zaufanie w tym kontekście można nazwać **globalnym średnim zaufaniem rekomendacyjnym**), równe:

$$t_{A\to A}^{c_2;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} t_{a_i\to a_j}^{c_2;m_l}}{n(n-1)}$$

- globalne średnie zaufanie w kontekście c_3 (które analogiczne jak zaufanie w tym kontekście można nazwać **globalnym średnim zaufaniem całkowitym**), równe:

$$t_{A\to A}^{c_3;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} t_{a_i\to a_j}^{c_3;m_l}}{n(n-1)}$$

W przypadku gdy istnieją w środowisku agenty złośliwe, globalne średnie zaufanie nie odzwierciedla prawidłowości funkcjonowania środowiska oraz wiarygodności systemu TRM. Z tego względu warto stworzyć bardziej szczegółowe miary w odniesieniu do różnych grup agentów i to właśnie ich używać do oceny funkcjonowania środowiska. W szczególności, istotne jest stworzenie miary globalnego średniego zaufania agentów rzetelnych do wszystkich agentów, agentów rzetelnych do agentów rzetelnych oraz agentów rzetelnych do agentów złośliwych.

Definicja 5.1.3.2. ***Globalne średnie zaufanie agentów rzetelnych do wszystkich agentów w kontekście c_k*** *to suma wartości zaufania w kontekście c_k pomiędzy agentem rzetelnym, a każdym innym agentem, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do wszystkich agentów ($A_B \to A$) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$t_{A_B\to A}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n} 1}$$

przy czym i, j jest takie, że: $a_i \in A_B, a_j \in A, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.

Definicja 5.1.3.3. ***Globalne średnie zaufanie agentów rzetelnych do agentów rzetelnych w kontekście c_k*** *to suma wartości zaufania w kontekście c_k pomiędzy każdą uporządkowaną parą agentów rzetelnych, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do agentów rzetelnych ($A_B \to A_B$) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$t_{A_B\to A_B}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n_B} 1}$$

przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_B, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.

Definicja 5.1.3.4. ***Globalne średnie zaufanie agentów rzetelnych do agentów złośliwych w kontekście c_k*** *to suma wartości zaufania w kontekście c_k każdego agenta rzetelnego do każdego agenta złośliwego, o ile pomiędzy tymi agentami zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do agentów złośliwych ($A_B \rightarrow A_M$) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$t_{A_B \rightarrow A_M}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B} \sum_{j=1}^{n_M} t_{a_i \rightarrow a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B} \sum_{j=1}^{n_M} 1}$$

przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_M$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.

Powyższą miarę można w łatwy sposób rozwinąć, tak aby w sytuacji gdy w środowisku istnieje więcej niż jedna grupa złośliwych agentów, poszczególne grupy traktować oddzielnie i stworzyć globalne zaufanie agentów rzetelnych do pierwszej grupy agentów złośliwych, do drugiej grupy agentów złośliwych, itd.

Analogiczne miary można zdefiniować w odniesieniu do reputacji, tj. można wyznaczyć globalną średnią reputację agentów, globalną średnią reputację agentów rzetelnych oraz globalną średnią reputację agentów złośliwych, zgodnie z poniższymi definicjami.

Definicja 5.1.3.5. ***Globalna średnia reputacja agentów w kontekście c_k*** *to suma wartości reputacji w kontekście c_k wszystkich agentów, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości reputacji. Globalna średnia reputacja wszystkich agentów (A) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$p_A^{c_k;m_l} = \frac{\sum_{j=1}^{n} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n} 1}$$

przy czym j jest takie, że $a_j \in A$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).

Definicja 5.1.3.6. ***Globalna średnia reputacja agentów rzetelnych w kontekście c_k*** *to suma wartości reputacji w kontekście c_k agentów rzetelnych, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości reputacji. Globalna średnia reputacja agentów rzetelnych (A_B) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$p_{A_B}^{c_k;m_l} = \frac{\sum_{j=1}^{n_B} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_B} 1}$$

przy czym j jest takie, że $a_j \in A_B$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).

Definicja 5.1.3.7. ***Globalna średnia reputacja agentów złośliwych w kontekście c_k*** *to suma wartości reputacji w kontekście c_k agentów złośliwych, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości reputacji. Globalna średnia reputacja agentów złośliwych (A_M) w kontekście c_k w chwili m_l jest wyrażona wzorem:*

$$p_{A_M}^{c_k;m_l} = \frac{\sum_{j=1}^{n_M} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_M} 1}$$

przy czym j jest takie, że $a_j \in A_M$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).

Miarę globalnej średniej reputacji agentów złośliwych w określonym kontekście można w łatwy sposób rozwinąć, tak aby w sytuacji gdy w środowisku istnieje więcej niż jedna grupa złośliwych agentów, poszczególne grupy traktować oddzielnie i stworzyć miarę globalnej reputacji pierwszej grupy agentów złośliwych, drugiej grupy agentów złośliwych, itd.

5.1.4. Miary średniego zaufania

Definicja 5.1.4. ***Średnie zaufanie do agenta a_j w kontekście c_k*** *to suma wartości zaufania w kontekście c_k każdego agenta do agenta a_j, o ile pomiędzy tymi agentami zaufanie jest określone, podzielona przez iloczyn maksymalnej wartości zaufania i liczby takich par agentów. Średnie zaufanie do agenta a_j w kontekście c_k w chwili m_l jest wyrażone wzorem:*

$$t_{A \to a_j}^{c_k;m_l} = \frac{\sum_{i=1}^{n} t_{a_i \to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n} 1}$$

przy czym i jest takie, że: $a_i \in A, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.

Uwaga 5.1.4.1. Średnie zaufanie do agenta można w praktyce utożsamiać z jego reputacją.

Uwaga 5.1.4.2. W ogólnym przypadku wartość średniego zaufania $t_{A\to a_j}^{c_k;m_l}$ nie musi należeć do zbioru możliwych wartości zaufania T.

5.1.5. Miary popularności i jakości agentów

Miary popularności agentów określają jak często z usług danej grupy agentów (lub konkretnego agenta) korzystają pozostałe agenty w środowisku, miary jakości wskazują z jaką średnią jakością dany agent lub grupa agentów świadczy swoje usługi.

Wprowadźmy następujące oznaczenia:

- $I^{P:a_j}$ – zbiór interakcji, w których agent a_j jest usługodawcą, $I^{P:a_j} \subseteq I$;
- $I^{R:a_i}$ – zbiór interakcji, w których agent a_i jest usługobiorcą, $I^{R:a_i} \subseteq I$;
- $I^{R:a_i,P:a_j}$ – zbiór interakcji, w których agent a_i jest usługobiorcą, a agent a_j jest usługodawcą, $I^{R:a_i,P:a_j} \subseteq I$, $I^{R:a_i,P:a_j} \subseteq I^{P:a_j}$, $I^{R:a_i,P:a_j} \subseteq I^{R:a_i}$;
- $i_k^{P:a_j}$ – k-ty element zbioru interakcji, w których agent a_j jest usługodawcą;
- $i_k^{R:a_i}$ – k-ty element zbioru interakcji, w których agent a_i jest usługobiorcą;
- $i_k^{R:a_i,P:a_j}$ – k-ty element zbioru interakcji, w których agent a_j jest usługodawcą a agent a_i jest usługobiorcą;
- $l_I^{P:a_j} = \left|I^{P:a_j}\right|$ – liczba interakcji, w których agent a_j jest usługodawcą;
- $l_I^{R:a_i} = |I^{P:a_i}|$ – liczba interakcji, w których agent a_i jest usługobiorcą;
- $l_I^{R:a_i,P:a_j} = \left|I^{R:a_i,P:a_j}\right|$ – liczba interakcji, w których agent a_j jest usługodawcą, a agent a_i usługobiorcą;
- $o_\Sigma^{P:a_j}$ – suma rzeczywistych wyników interakcji, w których agent a_j był usługodawcą

Na tej podstawie można zdefiniować następujące miary popularności agentów:

Definicja 5.1.5.1. ***Liczba interakcji agentów rzetelnych z agentami rzetelnymi*** *to suma liczby interakcji, w której zarówno usługodawca a_j oraz usługobiorca a_i byli rzetelni:*

$$l_I^{A_B,A_B} = \sum_{i,j:a_i,a_j\in A_B,i\neq j} l_I^{R:a_i,P:a_j}$$

Definicja 5.1.5.2. ***Liczba interakcji agentów rzetelnych z agentami złośliwymi*** *to suma liczby interakcji pomiędzy agentami, w której jeden z agentów był rzetelny, a drugi złośliwy:*

$$l_I^{A_B,A_M} = \sum_{i:a_i \in A_B, j:a_j \in A_M} l_I^{R:a_i,P:a_j} + \sum_{i:a_i \in A_M, j:a_j \in A_B} l_I^{R:a_i,P:a_j}$$

Warto zwrócić uwagę na fakt, że w skutecznie działającym środowisku liczba interakcji agentów rzetelnych z agentami rzetelnymi powinna być znacząco wyższa niż liczba interakcji agentów rzetelnych z agentami złośliwymi[55] i właśnie takiej ocenie służą powyższe dwie miary.

System TRM może bardzo mocno wpłynąć na nierównomierne obciążenie agentów w systemie (poniekąd taki jest jego cel). Problemem jednak może być sytuacja, w której jedynie pewna podgrupa agentów rzetelnych świadczy usługi, a pozostałe nie. Z tego względu nie tylko miary ujęte w definicjach 5.1.5.1. i 5.1.5.2. mają znaczenie, istotny jest także rozkład liczby interakcji poszczególnych agentów, będących usługodawcami, tj. $l_I^{P:a_j}$.

Definicja 5.1.5.3. ***Średni rzeczywisty wynik usług świadczonych przez agenta a_j,*** *który wyświadczył przynajmniej jedną usługę, to iloraz sumy rzeczywistych wyników interakcji i liczby interakcji, w których ten agent był usługodawcą:*

$$\forall_{a_j \in A_P, l_I^{P:a_j} > 0} \; o_E^{P:a_j} = \frac{o_\Sigma^{P:a_j}}{l_I^{P:a_j}}$$

Definicja 5.1.5.4. ***Średnia liczba interakcji, w których agenci byli usługodawcami $\overline{l_I^{P:A}}$,*** *to iloraz sumy liczba interakcji, w których agenci, którzy wyświadczyli przynajmniej jedną usługę byli usługodawcami i liczby takich agentów:*

$$\overline{l_I^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} l_I^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$$

Uwaga 5.1.5.1.: Średnia liczba interakcji, w których agenci byli usługodawcami $\overline{l_I^{P:A}}$ uwzględnia jedynie tych agentów, którzy dostarczyli przynajmniej jedną usługę.

[55] Łatwo zauważyć, że generowanie przez złośliwe agenty żądań świadczenia usług skierowanych do rzetelnych agentów także wpływa na zwiększenie liczby interakcji agentów rzetelnych z agentami złośliwymi, co w przypadku niektórych typów badań może nie być prawidłowe, dlatego w takim przypadku należałoby pominąć drugi składnik we wzorze określonym w definicji 5.1.5.2.

Definicja 5.1.5.5. Średnia rzeczywista jakość usług ($\overline{o_E^{P:A}}$) *jest średnią arytmetyczną średnich rzeczywistych wyników usług świadczonych przez agentów, którzy dostarczyli przynajmniej jedną usługę.*

$$\overline{o_E^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} o_E^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$$

W powyższych definicjach wszystkie usługi świadczone przez agentów zostały potraktowane łącznie, oczywiście można rozwinąć te miary w celu ich wyznaczenia oddzielnie dla każdej z usług.

5.1.6. Miary kosztu działania systemu TRM

Miary kosztu działania systemu TRM mogą określać jakie koszty są ponoszone m.in. przez agenty ze względu na funkcjonowanie systemu TRM. Takie miary pozwalałaby też zidentyfikować możliwe słabe punkty systemów TRM i wskazać potencjalne ataki na nie. Na przykład znaczna liczba wydanych rekomendacji, włącznie z koniecznością przeprowadzenia złożonych obliczeń, może sugerować, że dany system TRM może być podatny na atak DoS polegający na wysyłaniu znacznej liczby żądań wydawania rekomendacji. Zdefiniowanie takich miar wymagałoby jednak więcej danych na temat środowiska oraz doszczegółowienia jego modelu. Zagadnienie to nie dotyczy bezpośrednio działań złośliwych agentów i wykracza poza zakres rozprawy, wobec czego zostanie pominięte.

5.1.7. Idealny system TRM

Za publikacją [79] autora rozprawy i na podstawie zdefiniowanych miar można określić, że idealny system TRM to taki system, który pomimo obecności złośliwych agentów w środowisku i przeprowadzania przez nie ataków, pozwala na osiągnięcie następujących wartości poszczególnych miar zdefiniowanych w tym podrozdziale:

Miary efektywności

W przypadku każdego środowiska, im większą wartość mają poszczególne miary efektywności, tym lepiej ono funkcjonuje, dlatego poszczególne miary efektywności powinny być zbliżone do jedności:

$$E_q = 1$$

W przypadku środowiska bez zakłóceń, także: $E = 1$

$$\forall_n : E^{(n)} = 1$$

$$E^{ideal} = 1$$

$$E^{adv} = 1$$

Miary zysku systemu:

Zysk oraz zysk absolutny powinien być znacząco większy od 0. Teoretycznie można by wskazać, że wartości te powinny być bliskie jedności, ale może to się okazać niemożliwe do osiągnięcia nie ze względu na sam system TRM, ale na to, że nawet bez jego obecności w danym środowisku, złośliwe agenty nie będą w stanie maksymalnie zmniejszyć efektywności (do 0).

$$G \gg 0$$

$$G_A \gg 0$$

Oczywiście stosowanie jedynie miar efektywności oraz zysku systemu jest niewystarczające, bo strategicznym celem atakujących może nie być pogorszenie efektywności systemu. Dlatego istotna jest także analiza kolejnych parametrów.

Miary globalnego średniego zaufania i reputacji:

Wpływ złośliwych agentów może być mierzony za pomącą miar globalnego średniego zaufania i reputacji – im wyższe zaufanie do agentów rzetelnych i im niższe do agentów złośliwych, tym teoretycznie mniej złośliwych działań będą mogły podejmować złośliwe agenty. W przypadku gdy rozważane jest globalne średnie zaufanie lub globalna średnia reputacja w odniesieniu do wszystkich agentów w środowisku, to nawet w przypadku funkcjonowania idealnego systemu TRM, wartość tych parametrów będzie uzależniona od stosunku liczby złośliwych i rzetelnych agentów.

$t_{A\to A}^{c_k;m_l}$, $t_{A_B\to A}^{c_k;m_l}$ – różna w zależności od stosunku liczby agentów złośliwych i rzetelnych

$$t_{A_B\to A_B}^{c_k;m_l} = t_{max}$$

$$t_{A_B \to A_M}^{c_k;m_l} = t_{min}$$

$p_A^{c_k;m_l}$ – różna w zależności od stosunku liczby agentów złośliwych i rzetelnych

$$p_{A_B}^{c_k;m_l} = p_{max}$$

$$p_{A_M}^{c_k;m_l} = p_{min}$$

Miary średniego zaufania oraz reputacji

Średnie zaufanie do każdego agenta rzetelnego lub reputacja tego agenta powinny osiągać maksymalną wartość w idealnym systemie TRM, analogicznie średnie zaufanie do każdego agenta złośliwego lub reputacja tego agenta powinny osiągać minimalną możliwą wartość.

$$\forall_{a_j \in A_B}: t_{A \to a_j}^{c_k;m_l} = t_{max}$$

$$\forall_{a_j \in A_M}: t_{A \to a_j}^{c_k;m_l} = t_{min}$$

$$\forall_{a_j \in A_B}: p_{a_j}^{c_k;m_l} = p_{max}$$

$$\forall_{a_j \in A_M}: p_{a_j}^{c_k;m_l} = p_{min}$$

Miary popularności i jakości agentów

Agenty rzetelne powinny nawiązywać interakcje tylko z agentami rzetelnymi, w związku z tym:

$$l_I^{A_B,A_B} = l_I$$

$$l_I^{A_B,A_M} = 0$$

W przypadku braku zakłóceń, średni rzeczywisty wynik usług świadczonych przez rzetelnego agenta a_j, tj. $o_E^{P:a_j}$ powinien być równy średniej maksymalnych jakość usług, świadczonych przez tego agenta ważonych liczbą wyświadczonych poszczególnych usług. W przypadku gdy średni rzeczywisty wynik usług świadczonych przez tego agenta, tj. $o_E^{P:a_j}$ jest niższy, oznacza to, ze agent ten wykonywał ataki, czyli nie jest agentem rzetelnym. Podobnie, średnia rzeczywista jakość usług $\overline{o_E^{P:A}}$, dla agentów rzetelnych powinna być równa średniej maksymalnych jakość usług $q_{max}^{u_l}$, świadczonych przez te agenty ważonych liczbą wyświadczonych poszczególnych usług.

Średnia liczba interakcji, w których agenty rzetelne były usługodawcami powinna być znacznie większa od średniej liczby interakcji, w której agenty złośliwe były usługodawcami. Ta druga wartość powinna być maksymalnie bliska 0:

$$\overline{l_I^{P:A_B}} \gg \overline{l_I^{P:A_M}}, \qquad l_I^{P:A_M} \to 0$$

Średnia liczba interakcji, w których agenci byli usługodawcami $l_I^{P:\mathrm{A}}$ jest zależna od liczby agentów złośliwych i rzetelnych oraz ogólnej liczby interakcji.

Powyższe określenie, jakimi parametrami powinien cechować się idealny system TRM, pozwala na dokonywanie porównania efektywności różnych systemów TRM w sposób ilościowy.

5.2. RODZAJE BADAŃ

Badania odporności systemów TRM na ataki opierają się na zdefiniowaniu środowiska, systemu TRM oraz przeprowadzanych ataków, zgodnie z przedstawionymi modelami.

Badania mają za zadanie sprawdzić jak zachowuje się środowisko (z działającym systemem TRM) w przypadku stosowania przez agenty złośliwych działań. Przeprowadzane badania mogą brać pod uwagę zmiany w środowisku (np. w liczbie agentów, w liczbie agentów rzetelnych, świadczonych usługach, charakterystyce agentów), zmiany w funkcjonowaniu danego systemu TRM (w tym jego parametrów), czy zmiany w stosowanych atakach i ich parametrach. Dodatkowym aspektem badań, będącym częścią charakterystyki środowiska w myśl przedstawionego modelu, niemniej wymagającym oddzielnego potraktowania jest charakterystyka żądań agentów. Charakterystyka żądań została wydzielona ze środowiska, jako że w danym środowisku żądania mogą przebiegać w różny sposób.

Podsumowując, aspekty badania są następujące:

- środowisko,
- charakterystyka żądań,
- system TRM,
- atak.

Warto podkreślić, że zarówno w przypadku ataku jak i systemu TRM może nie chodzić jedynie o sam atak, czy system, ale także o parametry używane przez dany system TRM lub atak (np. w przypadku ataku oscylacji zachowania – istotny jest stosunek pomiędzy „czasem" rzetelnego zachowania a „czasem" nierzetelnego zachowania danego agenta).

Każdy z powyższych aspektów może być:

- **ustalony** (tzn. w każdym przeprowadzanym badaniu dany aspekt jest taki sam, np. przyjmujemy, że w środowisku nie zachodzą żadne zmiany w kolejnych badaniach);

- **zmienny** (tzn. w każdym przeprowadzanym badaniu charakterystyka danego aspektu jest inna, ale mogła zostać z góry określona przed badaniem, np. przeprowadzane są badania wpływu różnych, z góry określonych ataków), jeżeli określony aspekt jest zmienny, to dane badanie jest złożone z wielu badań składowych, a w każdym z nich dany aspekt jest inaczej określony;
- **uogólniony** (charakterystyka danego aspektu jest określana w ramach badania, a przebieg samego badania ma wpływ na ten aspekt, czyli że nie mógł on zostać określony przed rozpoczęciem badania, np. badanie ma na celu wyznaczyć takie zachowanie złośliwych agentów, które doprowadzi do np. największego spadku efektywności środowiska, tj. ma za zadanie zidentyfikować potencjalnie najbardziej efektywny atak).

Najprostszym rodzajem badania jest badanie, w którym każdy z aspektów jest ustalony, tzn. jest to badanie określonego niezmiennego środowiska, z ewentualnie działającym ustalonym systemem TRM i jego parametrami, z ustaloną charakterystyką żądań i określonymi działaniami złośliwych agentów. W takim przypadku wystarczające jest wyznaczenie poszczególnych miar, zdefiniowanych w poprzednim podrozdziale. Może być przydatne także przeprowadzenie takiego badania wielokrotnie z uwagi na często występujące elementy losowe w danym środowisku.

Z drugiej strony, interesującym, aczkolwiek teoretycznym rodzajem badania, byłoby badanie, w którym wszystkie aspekty są uogólnione. Jeżeli byłoby możliwe przeprowadzenie takiego badania, jego wynik wskazałby na optymalny system TRM do użycia w każdym środowisku, mimo prowadzenia optymalnych działań złośliwych agentów. Rzecz jasna konstrukcja metody przeprowadzenia takiego badania jest obiektywnie trudna.

Już opracowanie metody przeprowadzenia badania, w którym jeden z aspektów jest uogólniony, a pozostałe ustalone jest wysoce nietrywialne. Rozważmy takie badania, w których poszczególne aspekty są uogólnione:

- **Badanie z uogólnionym środowiskiem** pozwoliłoby znaleźć odpowiedź na pytanie jakie musiałoby być środowisko żeby dany atak na dany system TRM z określoną charakterystyką żądań przyniósł największą degradację efektywności. Wydaje się, że wynik takiego badania byłby mniej interesujący, ze względu na to, że raczej istnieje potrzeba dopasowania systemu TRM i jego parametrów po to żeby osiągnąć najwyższą efektywność, natomiast zwykle nie ma możliwość dostosowywania środowiska.

- **Badanie z uogólnioną charakterystyką żądań** pozwoliłoby wskazać taki przebieg żądań, który przy innych czynnikach niezmienionych doprowadziłby do osiągnięcia najwyższej lub najniższej wartości efektywności lub innych miar. Podobnie jak w przypadku środowiska, wynik takiego badania wydaje się być mało interesujący, gdyż trudno spodziewać się aby na poziomie rozproszonego środowiska możliwe byłoby sterowanie napływającymi żądaniami, jest to wszakże charakterystyka wejściowa.
- **Badanie z uogólnionym systemem TRM** pozwala określić cechy, które musi mieć system, aby zapobiec danemu atakowi w określonym środowisku. Wynik badań tego rodzaju jest bardzo interesujący dla twórców systemów TRM. Warto jednak podkreślić, że zapobieganie danemu atakowi może doprowadzić do istnienia podatności na inny rodzaj ataku. Opracowanie metody badania z uogólnionym systemem TRM to bardzo interesujący temat badawczy, ale wykraczający poza zakres niniejszej pracy, gdyż jej ideą jest ocena odporności systemów TRM oraz ewentualne wskazanie dla danego systemu, najbardziej efektywnego ataku.
- **Badanie z uogólnionym atakiem** ma na celu wskazanie działań atakujących, które w danym środowisku będą najskuteczniejsze (np. zminimalizują efektywność środowiska). Ten typ badania został zaprezentowany w punkcie 5.3.7. (opisującym proces tworzenia tzw. ataku dopasowanego) oraz w punkcie 5.3.8 i rozdziale 5.4 (które opisują tzw. metodę MEAEM).

Istotne jest podkreślenie różnicy pomiędzy tym, że dany aspekt jest zmienny a uogólniony. W przypadku badania uogólnionego, powinny zostać rozważone wszystkie możliwe przypadki charakterystyki danego aspektu (których liczba jest zwykle nieskończona) lub te przypadki, które są w stanie wygenerować skrajne wartości poszczególnych zdefiniowanych miar (co oczywiste identyfikacja takich przypadków jest wysoce nietrywialna). W przypadku badania, w którym dany aspekt jest zmienny, następuje wiele badań składowych, w których ten aspekt jest ustalony. Warto zwrócić uwagę, że najczęściej stosowana w literaturze analiza polegająca na badaniu systemu (jednego lub kilku) w stosunku do znanego zbioru ataków to nie badanie z uogólnionym atakiem, lecz badanie ze zmiennym atakiem.

Dane badanie (lub badanie składowe) może składać się z wielu przebiegów badania/symulacji, a wyniki w kolejnych przebiegach symulacji mogą się różnić od siebie wskutek istniejącego w systemie TRM lub środowisku elementu losowego (np. system TRM uwzględnia zmienną losową w funkcji wyboru usługodawcy lub występują zakłócenia

w środowisku). Taka zmienność wyników w poszczególnych przebiegach może ujawniać interesujące właściwości danego systemu TRM lub środowiska.

5.3. PROPOZYCJE BADAŃ

Zakładając, że jest dane środowisko (określone dokładnie, tj. opisane przez przedstawiony model lub określone jedynie poprzez podanie wybranych jego specyficznych aspektów) i system TRM, dla którego może istnieć konieczność wyboru pewnych parametrów, celem jest sprawdzenie jakości takiego systemu i jego odporności na ataki. Aby zrealizować ten cel warto rozważyć przeprowadzenie następujących badań.

5.3.1. Wstępna ocena podatności na ataki

Badanie polega na przeprowadzeniu analizy funkcjonowania systemu TRM i środowiska i służy wstępnej ocenie podatności na ataki z uwzględnieniem specyfiki środowiska, systemu TRM i potencjalnych możliwości atakującego.

Cele wstępnej oceny są następujące:

- wykluczenie niektórych ataków (np. ze względu na to, że atakujący nie będą posiadać wystarczającej wiedzy koniecznej do przeprowadzenia ataku, lub ze względu na specyfikę systemu lub środowiska, atak może okazać się niemożliwy);
- wykluczenie miar wiarygodności, które nie mają zastosowania (np. miar dotyczących globalnej średniej reputacji w przypadku gdy system TRM nie korzysta z pojęcia reputacji).

5.3.2. Badanie środowiska bez systemu TRM

Badanie służy sprawdzeniu jak funkcjonowałoby dane środowisko bez stosowania systemu TRM. W przypadku takiego badania, uzasadnione jest analizowanie jedynie niektórych miar wiarygodności (w szczególności miar efektywności).

Założenia:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- brak systemu TRM,

- ustalony lub zmienny atak, ustalone parametry ataku (domyślne).

W przypadku istnienia elementów losowych, badanie powinno być przeprowadzone wielokrotnie.

5.3.3. Badanie reakcji ustalonego systemu TRM na ustalone ataki

Głównym celem badania jest zgrubna jakościowa ocena, czy system TRM jest podatny na dany rodzaj ataku, czy też nie.

Założenia:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- ustalony system TRM, ustalone parametry systemu TRM (domyślne),
- ustalony lub zmienny atak, ustalone parametry ataku (domyślne).

W przypadku istnienia elementów losowych, badanie powinno być przeprowadzone wielokrotnie.

5.3.4. Badanie wpływu wartości parametrów systemu TRM

Badanie służy porównaniu wartości poszczególnych miar efektywności w przypadku stosowania różnych parametrów systemu TRM i próbie dokonania ich optymalizacji.

Założenia:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- ustalony system TRM, zmienne parametry systemu TRM,
- ustalony lub zmienny atak, ustalone parametry ataku (domyślne).

Każdy zestaw parametrów systemów TRM wygeneruje kolejne badanie składowe. W przypadku istnienia elementów losowych, każde badanie składowe powinno być przeprowadzone wielokrotnie.

5.3.5. Badanie wpływu doboru parametrów ataków

Badanie służy ocenie danego systemu TRM w przypadku stosowania znanych ataków, ale które przyjmują specyficzne parametry.

Założenia:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- ustalony system TRM, ustalone parametry systemu TRM (domyślne),
- ustalony lub zmienny atak, zmienne parametry ataku.

Każdy zestaw parametrów ataku wygeneruje kolejne badanie składowe. W przypadku istnienia elementów losowych, każde badanie składowe powinno być przeprowadzone wielokrotnie.

5.3.6. Badanie wpływu parametrów środowiska

Badanie ma na celu rozstrzygnięcie jak zmiana charakterystyki środowiska wpływa na wiarygodność stosowanego systemu TRM.

Założenia:

- zmienne parametry środowiska, np. zmienna liczba agentów w środowisku, zmienna liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań
- ustalony system TRM, ustalone parametry systemu TRM (domyślne),
- ustalony lub zmienny atak, ustalone parametry ataku (domyślne).

Każdy zestaw parametrów środowiska wygeneruje kolejne badanie składowe. W przypadku istnienia elementów losowych, każde badanie składowe powinno być przeprowadzone wielokrotnie.

Przykładowe badania mogą mieć na celu sprawdzenie:

- czy wielkość środowiska (liczba agentów) ma wpływ na uzyskane wyniki,
- w przypadku środowiska heterogenicznego (np. usługa przekazywania pakietów), czy istnieje wpływ „lokalizacji” lub parametrów agentów złośliwych na uzyskiwane wyniki.

5.3.7. Tworzenie i badanie ataku dopasowanego

W przeciwieństwie do innych badań zaproponowanych w niniejszym podrozdziale, badanie to będzie możliwe jedynie w przypadku niektórych środowisk i systemów TRM. Jego założeniem jest to, że pogłębiona analiza specyficznych właściwości systemu TRM i jego

parametrów może doprowadzić do opracowania kreatywnego sposobu na jego zmanipulowanie.

Założenia badania:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- ustalony system TRM, ustalone parametry systemu TRM (domyślne),
- próba stworzenia ataku, który wykorzysta specyficzne właściwości systemu TRM i jego parametrów.

Stworzenie takiego ataku i przeprowadzenie badań odporności systemu, będzie przebiegało odmiennie w odniesieniu do każdego systemu TRM (a także konkretnego środowiska). Przykład takiego ataku skutecznego przeciwko określonemu środowisku oraz systemowi RefTRM został zaprezentowany w punkcie 6.3.7. Ogólnie można stwierdzić, że przy próbie stworzenia ataku dopasowanego należy wziąć następujące właściwości systemu TRM i środowiska:

- zakres zmienności wartości jakości usług,
- zakres zmienności wartości rekomendacji,
- analizę wszelkich wartości progowych używanych przy obliczaniu wartości zaufania lub ocenie jakości usługi,
- wyniki poprzednich badań, które mogą ujawnić interesujące właściwości systemu.

5.3.8. Badanie z uogólnionym atakiem

Badanie z uogólnionym atakiem ma na celu identyfikację najbardziej efektywnego ataku w danym środowisku stosującym określony system TRM. Badanie to może zostać przeprowadzone zgodnie z metodą MEAEM, która zostało opisana szczegółowo w kolejnym podrozdziale. Warto zauważyć, że ta metoda pozwala na identyfikację nowych ataków przeciwko systemom TRM.

Założenia:

- ustalone środowisko, w tym ustalona liczba agentów w środowisku, ustalona liczba agentów złośliwych,
- ustalona lub zmienna charakterystyka żądań,
- ustalony system TRM, ustalone parametry systemu TRM (domyślne),
- uogólniony atak.

5.4. METODA MEAEM – BADANIE Z UOGÓLNIONYM ATAKIEM

Metoda oceny najbardziej efektywnego ataku (MEAEM – ang. „Most Effective Attack Evaluation Method") została po raz pierwszy opisana w artykule [83] autora niniejszej rozprawy i *„polega na próbie analizy możliwych przypadków zachowań agentów i na tej podstawie określenia jakie działania powinny zostać podjęte przez złośliwe agenty w celu zmniejszenia efektywności funkcjonowania środowiska. Istotną różnicą tej metody w stosunku do większości badań symulacyjnych jest to, iż pozwala ona na podjęcie decyzji przez złośliwe agenty na bazie analizy funkcjonowania systemu TRM i aktualnych warunków panujących w środowisku. Metoda ta, w przeciwieństwie do innych badań symulacyjnych, nie zakłada a priori sposobu działania złośliwych agentów, a skupia się na próbie rozważenia możliwych działań atakujących i wyboru takich, które przynoszą im największą korzyść. Jest to szczególnie istotne ze względu na to, że działania atakujących mogą składać się z wielu z opisanych ataków, tzn. atakujący mogą stosować strategię mieszaną*[56]*. W przypadku krańcowym inny atak elementarny może być stosowany przy każdej interakcji (i przy każdym wydaniu rekomendacji). Problem jest najbardziej skomplikowany gdy rozważamy wielu atakujących, którzy współpracują ze sobą (jest to idea kooperacyjnych ataków inteligentnych* [69], [108]*). Dana strategia może wykazywać cechy strategii mieszanej zarówno w sensie czasowym (agenty stosują różne ataki w różnym czasie) i w sensie przestrzennym (różne złośliwe agenty – wszystkie współpracujące ze sobą – stosują różne ataki w tym samym czasie). Warto podkreślić, że postępowanie zgodnie ze strategią mieszaną może skutkować wyższą skutecznością działań atakujących niż w wypadku innych strategii (składających się z wielu ataków).*

Metoda MEAEM stanowi próbę odpowiedzi na pytanie czy może istnieć taki sposób postępowania złośliwych agentów, który istotnie zakłóci funkcjonowanie danego systemu TRM. Metoda ta stanowi odejście od analizy znanych ataków. Dzięki temu możliwe będzie zmniejszenie zagrożenia polegającego na tym, że zostanie opracowany nowy atak, który nie został uwzględniony na etapie analizy systemu. Takie podejście potencjalnie umożliwia znalezienie ataku, który będzie skuteczniejszy od znanych ataków, a w pewnych sytuacjach znalezienie najgorszego przypadku z perspektywy funkcjonowania środowiska z zaimplementowanym systemem TRM.

[56] Pojęcie strategia mieszana w tym kontekście nie jest tożsame z pojęciem strategii mieszanej w ramach teorii gier. Warto podkreślić, że przedstawione rozumowanie nie ogranicza się do klasy systemów zarządzania zaufaniem i reputacją opartych o teorię gier, ale jest właściwe dla wszystkich systemów wykorzystujących pojęcie zaufania lub reputacji.

Mimo tego, że idea wykonania w ten sposób ewaluacji systemu TRM jest prosta, to jej praktyczna realizacja wiąże się z następującymi problemami:

- *jak mierzyć korzyść dla atakujących i czy będzie to możliwe niezależnie od szczegółowego celu atakujących (np. czy większą korzyścią jest zwiększenie reputacji atakujących lub zaufania do nich, czy chwilowe zmniejszenie efektywności środowiska),*
- *jakie decyzje (akcje) mogą wykonywać atakujący,*
- *jak ograniczyć złożoność obliczeniową rozwiązania (rozważenie wszystkich przypadków już dla niewielkich środowisk i kilku interakcji staje się zadaniem złożonym obliczeniowo),*
- *jak stwierdzić czy sekwencja zachowań atakujących, która zostanie zaproponowana przez tę metodę będzie możliwa do realizacji w praktyce (tzn. czy w rzeczywistych warunkach, złośliwe agenty będą dysponowały wystarczającymi informacjami umożliwiającymi podjęcie dokładnie tych samych decyzji).*

Problemem dodatkowo komplikującym wykonanie takiej analizy jest brak determinizmu działań agentów, rozumiany wielorako. Po pierwsze, agenty wybierając agenta do interakcji, w przypadku wielu systemów TRM, postępują niedeterministycznie (tzn. np. wybierają agenta spośród tych o najwyższej reputacji lub najwyższym zaufaniu, ale z pewnym prawdopodobieństwem uzależnionym od reputacji danego agenta lub zaufania do danego agenta). Po drugie, jakakolwiek rozważana zmiana zachowania agenta podczas danej interakcji, spowoduje, że zmienią się warunki w czasie trwania wszystkich kolejnych interakcji w stosunku do warunków rozważanych wcześniej. Istnieje także trzeci rodzaj braku determinizmu (ale o stosunkowo mniejszym znaczeniu). W rzeczywistych typach środowisk, w których są stosowane systemy TRM, występują wszakże niedeterministyczne zakłócenia, niewynikające z zachowania agentów, co dodatkowo utrudnia analizę."

Analiza MEAEM może być wykonana przy założeniu stałych:

- parametrów środowiska (liczby agentów w sieci i jej topologii, oraz charakterystyki interakcji),
- liczby złośliwych agentów,
- liczby interakcji (lub czasu działania środowiska),
- sekwencji i charakterystyki żądań usług, tzn. konieczne jest określenie, które agenty, kiedy i jakiej usługi żądają,
- wykorzystywanego systemu TRM.

Wedle najlepszej wiedzy autora metoda analogiczna do MEAEM nie została zaprezentowana dotychczas w literaturze, jednak została natomiast zidentyfikowana potrzeba dokonania takiej analizy (aczkolwiek wyrażona nie wprost), m in. w pracach: [69], [108].

Najprostszym sposobem przeprowadzania takiej analizy byłaby próba rozważenia wszystkich możliwych zachowań jednostkowych złośliwych agentów wedle modelu ataku w oparciu o zachowanie jednostkowe (określonego w punkcie 4.3.1) dla każdej interakcji. Jak zauważył autor rozprawy w publikacji [106]: *„na wybór optymalnego zachowania agenta w interakcji k ma wpływ zachowanie wszystkich agentów w interakcjach 1, …, k-1 (zachowania agentów w poszczególnych interakcjach są od siebie niezależne, ale mają wpływ na to jakie zachowanie będzie najlepsze z perspektywy złośliwych agentów). Inaczej rzecz ujmując, agent złośliwy podejmując daną decyzję w czasie konkretnej interakcji nie jest w stanie stwierdzić, która z nich będzie optymalna w perspektywie osiągnięcia celu przez atakujących. W związku z tym istnieje konieczność zasymulowania wszystkich decyzji w każdej interakcji i dopiero po zakończeniu analizy wybranie tych, które zapewniły najlepszy wynik dla atakujących."* Nietrudno zauważyć, że nawet przy założeniach upraszczających, wykonanie analizy wszystkich możliwych zachowań agentów jest obliczeniowo niewykonalne we wszystkich typowych zastosowaniach. W szczególności, gdy agenty mogą wydawać rekomendację zawierającą wartość z nieskończonego zbioru wartości możliwych rekomendacji lub świadczyć usługę z jakością o wartości z nieskończonego zbioru, to w takiej sytuacji rozważenie wszystkich możliwych zachowań agenta już w pojedynczej interakcji jest niemożliwe, ponieważ liczba takich zachowań jest nieskończona. Z tego powodu konieczna jest próba stworzenia algorytmu heurystycznego umożliwiającego osiągnięcie celu atakujących.

Algorytm heurystyczny ma za zadanie określenie najlepszych decyzji złośliwych agentów, takich jak:

- jak wybrać agenta do interakcji,
- jaką wydać rekomendację,
- jaką jakość usługi dostarczyć,
- jakiego identyfikatora używać w każdym z powyżej określonych etapów.

Decyzje te mają doprowadzić do osiągnięcia przez złośliwe agenty maksymalnego zysku przy danych warunkach panujących w środowisku.

5.4.1. Konsekwencje działań atakujących

Rozważmy możliwe konsekwencje działań atakujących w odniesieniu do poszczególnych zachowań jednostkowych, zgodnie z modelem ataku przedstawionym w podrozdziale 4.3.1. Konsekwencje te zostały także wskazane we wcześniejszej publikacji autora rozprawy [79]. Przyjmijmy następujące oznaczenia:

- (+) – identyfikuje konsekwencje korzystne dla atakujących,
- (-) – identyfikuje konsekwencje niekorzystne dla atakujących.

Manipulacja jakością dostarczanej usługi (zachowanie $\boldsymbol{Q^{-}}$) może pociągać za sobą:

- zmniejszenie efektywności środowiska: (+);
- zmniejszenie zaufania do agenta świadczącego usługę: (-).

Manipulacja rekomendacjami (zachowanie $\boldsymbol{R^{\pm}}$ lub $\boldsymbol{R^{0}}$) może pociągać za sobą:

- zmniejszenie zaufania do agentów rzetelnych (co może prowadzić do zmniejszenia efektywności): (+);
- zwiększenie zaufania do agentów złośliwych (co może prowadzić do zmniejszenia efektywności): (+);
- zmniejszenie zaufania do agenta wydającego rekomendację (-);
- sklasyfikowanie przez agenta otrzymującego rekomendację, agenta rzetelnego jako złośliwy lub na odwrót: (+).

Manipulacja tożsamością (zachowanie $\boldsymbol{I^{\times}}$ lub $\boldsymbol{I^{\neq}}$) może pociągać za sobą:

- zwiększenie oddziaływania agentów złośliwych na wartość zaufania: (+);
- bezpośrednie zmniejszenie zaufania do rzetelnego agenta poprzez podszycie się pod niego: (+);
- bezpośrednie zwiększenie własnej reputacji poprzez obranie nowej tożsamości: (+);
- konieczność wykonywania większej liczby obliczeń i wysyłania większej liczby rekomendacji: (-).

Metoda MEAEM przyjmuje, że złośliwe agenty są usługodawcami i nie wykonują ataków związanych z nieprawidłowym wyborem agentów do interakcji w przypadku gdy pełniłyby rolę usługobiorców. Wobec tego nie jest rozważana manipulacja wyborem usługodawcy (zachowanie $\boldsymbol{S^{\neq}}$).

5.4.2. Cele atakujących

Celem atakujących może być np.:

- minimalizacja efektywności środowiska E (np. średniej efektywności podczas całego działania środowiska),
- minimalizacja zaufania do agentów rzetelnych,
- maksymalizacja zaufania do agentów złośliwych,
- maksymalizacja liczby interakcji ze złośliwymi agentami $l_{:M}$[57].

Warto zauważyć, że optymalizacja działań atakujących nie powinna polegać na minimalizacji chwilowej efektywności środowiska, ponieważ chwilowy spadek efektywności może iść w parze ze znacznym spadkiem reputacji lub zaufania, co następnie może przełożyć się na brak możliwości wykonania kolejnych ataków, konieczne jest więc uwzględnienie szerszej perspektywy korzyści atakujących.

Jak zauważono wcześniej, nawet przy jednoznacznym określeniu celu atakujących, nie jest możliwe przeanalizowanie wszystkich możliwych działań atakujących w każdym momencie działania środowiska i z uwzględnieniem wszystkich czynników wpływających na te decyzje. Zgodnie z publikacją [83] autora rozprawy: *„ze względu na brak możliwości zastosowania algorytmu naiwnego, konieczne jest stworzenie algorytmu heurystycznego. Idea tego algorytmu opiera się na tym, aby podczas każdej interakcji złośliwe agenty zbliżały się do osiągnięcia zakładanego celu, czyli aby uzyskiwały korzyść ze złośliwego działania. Warto podkreślić, że cel atakujących nie musi być tożsamy z korzyścią atakujących podczas danej interakcji. Dla przykładu: w przypadku gdy celem atakujących jest minimalizacja efektywności środowiska podczas całego analizowanego okresu działania, optymalizacja działań atakujących nie powinna polegać tylko na minimalizacji chwilowej efektywności. Konieczne jest więc m.in. uwzględnienie przy rozważaniu korzyści atakujących ze złośliwego działania, m.in. wartości reputacji lub zaufania złośliwych agentów.”*

5.4.3. Przebieg metody

Przed rozpoczęciem analizy należy rozstrzygnąć jakie zachowania jednostkowe są możliwe, w szczególności dotyczy to ataku polegającego na manipulowaniu tożsamością. Warto zwrócić uwagę, że metoda polega na tym, że wybierana jest taka sekwencja działań atakujących (w kontekście określonych wcześniej zachowań jednostkowych), która pozwoli na

[57] Ten cel jest tożsamy z działaniami samolubnymi, w przypadku gdy za świadczenie usług, agenty otrzymują wynagrodzenie.

maksymalizację zysku atakujących (wyjaśnienie w dalszej części podrozdziału), a dopiero później zachodzi potrzeba ostatecznego rozstrzygnięcia, czy w rzeczywistych warunkach agenty miałyby możliwość przeprowadzenia ataku w taki sposób (np. czy posiadały wystarczające informacje konieczne do jego przeprowadzenia). Na cele dalszej analizy przyjęto założenie, że atakujący mają dostęp do wszystkich informacji.

Metoda MEAEM polega na określeniu funkcji zysku atakujących, zasymulowaniu w każdej interakcji możliwych działań złośliwych agentów, na tej podstawie wybraniu zestawu najlepszych decyzji z punktu widzenia zysku atakujących, a następnie na przejściu do analizy kolejnej interakcji. W przypadku wydawania rekomendacji a priori (przed wyborem usługodawcy i skorzystaniem z usługi przez agenta)[58], podczas każdej interakcji wykonywane są następujące kroki:

1. Określenie wartości efektywności chwilowej środowiska oraz wartości zaufania agentów rzetelnych do poszczególnych agentów lub globalnego średniego zaufania agentów rzetelnych do agentów rzetelnych i do agentów złośliwych w poszczególnych kontekstach używanych w systemie.
2. Przyjęcie wartości rekomendacji (wydawanych przez złośliwe agenty) oraz wartości jakości usług (świadczonych przez złośliwe agenty) z określonego zbioru (dalej nazywanych zestawem decyzji); dla każdego zestawu decyzji:
    a. wydanie rekomendacji przez złośliwe agenty dla rzetelnych agentów, o agentach rzetelnych oraz o agentach złośliwych;
    b. określenie prawdopodobieństwa wyboru jako usługodawcy wszystkich agentów, którzy mogą świadczyć daną usługę, lub prawdopodobieństwa wyboru jako usługodawcy dowolnego agenta ze zbioru agentów rzetelnych oraz złośliwych;
    c. zasymulowanie wyboru jako usługodawcy wszystkich agentów, którzy mogą świadczyć daną usługę, lub dowolnego agenta ze zbioru agentów rzetelnych oraz złośliwych; dla każdego przypadku:
        i. w przypadku gdy został wybrany agent złośliwy, świadczenie usługi o jakości określonej w zestawie decyzji, w przypadku gdy wybrany został agent rzetelny – świadczenie usługi zgodnie z jego charakterystyką (wtedy jest to niezależne od atakujących);
        ii. określenie wartości efektywności chwilowej środowiska oraz wartości zaufania agentów rzetelnych do poszczególnych agentów lub globalnego średniego zaufania

[58] Przypadek odwrotny tzn. kiedy rekomendacje (interakcji) wydawane są po interakcji może zostać przeanalizowany w analogiczny sposób.

agentów rzetelnych do agentów rzetelnych i do agentów złośliwych w poszczególnych kontekstach używanych w systemie;

iii. obliczenie zmiany wartości powyższych parametrów;

iv. analiza kolejnego przypadku (wyboru innego agenta jako usługodawcy) – realizacja punktu c, o ile nie wszystkie przypadki zostały uprzednio przeanalizowane;

d. realizacja punktu 2. dla kolejnego zestawu decyzji, jeżeli nie wszystkie zostały uprzednio przeanalizowane.

3. Wybór takiego zestawu decyzji, który zapewnia maksymalizację wartości funkcji zysku atakujących.
4. Wydanie rekomendacji przez wszystkie agenty złośliwe, zgodnie z wybranym zestawem decyzji.
5. Wybór agenta usługodawcy (krok niezależny od atakujących).
6. Interakcja:
   a. jeżeli został wybrany agent rzetelny – brak konieczności podejmowania decyzji (krok niezależny od atakujących);
   b. jeżeli został wybrany agent nierzetelny – ponowne przeprowadzenie symulacji pozwalającej na wybór jakości usługi, która zapewni maksymalizację wartości funkcji zysku atakujących[59] lub świadczenie usługi o wartości jakości zgodnie z zestawem decyzji określonym w punkcie 3.
7. Analiza kolejnej interakcji.

Metoda opiera się na wielokrotnie powtarzanej analizie możliwych działań złośliwych agentów związanych z jedną interakcją. Analiza działań w ramach różnych interakcji dokonywana jest niezależnie. Ponieważ decyzje w ramach każdej interakcji podejmowane są niezależnie od innych interakcji, mamy do czynienia z heurystyką. Warto zauważyć, że takie podejście, mimo że uzasadnione z uwagi na konieczność ograniczenia liczby możliwych przypadków do przeanalizowania, to może nie być skuteczne (tzn. może nie wygenerować optymalnego sposobu postępowania atakujących). W szczególności dotyczy to tych systemów, w których obowiązywanie danej rekomendacji wykracza poza podjęcie decyzji dotyczącej

[59] Przeprowadzenie ponownej analizy, jaką wartość jakości usługi zapewnić, może być konieczne, szczególnie w przypadku gdy podczas analizy zestawu decyzji, dokonywanej w punkcie drugim, nie wzięto pod uwagę wszystkich możliwych decyzji dokonywanych przez rzetelne agenty, niezależnie od złośliwych agentów, np. decyzji dotyczącej wyboru usługodawcy. W tamtym momencie złośliwe agenty mogły jedynie szacować prawdopodobieństwo wyboru danego agenta jako usługodawcy, w tym momencie wiedzą już, który agent został wybrany, co może mieć wpływ na optymalną dla atakującego decyzję w odniesieniu do jakości świadczonej usługi.

jednej interakcji (np. rekomendacja jest wydawana po interakcji i wszystkie agenty mogą z niej korzystać do końca działania środowiska). Warto podkreślić, że celem atakujących może być zmniejszenie efektywności podczas całego funkcjonowania środowiska. Zaufanie lub reputacja jest skorelowane z częstotliwością wyboru agenta jako usługodawcy. W związku z tym agenty dążą do: zmniejszenia efektywności, zwiększenia zaufania agentów rzetelnych do złośliwych oraz zmniejszenie zaufania agentów rzetelnych do innych agentów rzetelnych.

5.4.4. Analiza możliwych zachowań atakujących odnośnie świadczenia usługi i wydawania rekomendacji

Na mocy punktu 4.3.1. złośliwy agent może podjąć różne decyzje odnośnie wydawanej rekomendacji, może w szczególności dokonać jej zaniżania lub zawyżania. W ramach metody MEAEM konieczne jest przyjęcie skończonego zbioru wartości rekomendacji, którą może zastosować złośliwy agent. Wydając rekomendacje o innych złośliwych agentach oraz o rzetelnych agentach, złośliwy agent może stosować odmienne zbiory tych wartości rekomendacji (prawdopodobnie strategicznie będzie dążył do zaniżania rekomendacji dotyczących rzetelnych agentów oraz zawyżania rekomendacji dotyczących złośliwych agentów, ale nie zawsze tak musi być). Podobnie, na mocy punktu 4.3.1., złośliwy agent może świadczyć usługę o maksymalnej jakości zgodnie ze swoją charakterystyką, lub świadczyć usługę o niższej jakości, bądź nie świadczyć w ogóle żądanej usługi. Metoda MEAEM umożliwia dokonanie analizy działań atakujących w oparciu o skończone zbiory wartości jakości świadczonych usług i rekomendacji wydawanych przez złośliwe agenty. Wobec tego w ramach metody MEAEM rozważane są:

- rekomendacje, które mogą zostać dostarczone dla rzetelnych agentów o złośliwych agentach: $r_i^{A_M} \in R_{A_M} \subset R$, gdzie R_{A_M} – skończony $k_{R_{A_M}}$-elementowy zbiór wartości rekomendacji o złośliwych agentach w metodzie MEAEM, $\left|R_{A_M}\right| = k_{R_{A_M}}$, będący podzbiorem zbioru wartości rekomendacji R, określonego w punkcie 4.2.3,
- rekomendacje, które mogą zostać dostarczone dla rzetelnych agentów o rzetelnych agentach: $r_i^{A_B} \in R_{A_B} \subset R$, gdzie R_{A_B} – skończony $k_{R_{A_B}}$-elementowy zbiór wartości rekomendacji o rzetelnych agentach w metodzie MEAEM, $\left|R_{A_B}\right| = k_{R_{A_B}}$, będący podzbiorem zbioru wartości rekomendacji R, określonego w punkcie 4.2.3.
- jakości usług, które mogą zostać wyświadczone dla agentów: $q_i \in Q_M \subset Q$, gdzie Q_M – skończony k_{Q_M}-elementowy zbiór jakości usług świadczonych przez złośliwych

agentów, rozważanych w metodzie MEAEM, $|Q_M| = k_{Q_M}$, będący podzbiorem zbioru jakości usług Q, określonego w podrozdziale 4.1.

Dokonajmy oszacowania liczby sposobów, na ile złośliwe agenty mogłyby wydawać rekomendacje dotyczące innych agentów. Rozważmy skrajny przypadek, że w każdej interakcji agent wydaje rekomendacje o każdym innym agencie. Wtedy, jeżeli każdy złośliwy agent wydawałby rekomendację o każdym agencie rzetelnym niezależnie, to w czasie jednej interakcji byłby w stanie wydać rekomendacje na temat rzetelnych agentów na $\left(k_{R_{A_B}}\right)^{n_B-1}$ sposobów oraz na temat innych złośliwych agentów na $\left(k_{R_{A_M}}\right)^{n_M-1}$ sposobów (zakładamy, że agent nie żąda i nie wydaje rekomendacji o sobie samym, wszakże jeden z agentów rzetelnych żąda usługi i rekomendacji, a agenci złośliwi także nie wydają rekomendacji o sobie samym). Czyli jeden złośliwy agent mógłby wydać rekomendacje na $\left(k_{R_{A_B}}\right)^{n_B-1} * \left(k_{R_{A_M}}\right)^{n_M-1}$ sposobów, a wszystkie złośliwe agenty mogłyby wydać rekomendacje na $\left(\left(k_{R_{A_B}}\right)^{n_B-1} * \left(k_{R_{A_M}}\right)^{n_M-1}\right)^{n_M}$ sposobów. Liczba sposobów na ile mogą zostać wydane same rekomendacje uniemożliwia przeanalizowanie ich wszystkich nawet dla niewielkiej liczby agentów złośliwych i rzetelnych oraz nielicznych zbiorów rekomendacji R_{A_M} i R_{A_B}. Dodatkowo, jeżeli usługobiorca wybierze złośliwego agenta jako usługodawcę to będzie on mógł świadczyć usługi k_{Q_M} sposobów. W przypadku gdy złośliwe agenty nie są w stanie a priori rozstrzygnąć, który agent zostanie wybrany jako usługodawca, muszą zostać wzięte pod uwagę wszystkie możliwości, czyli muszą rozważyć możliwość świadczenia usług na $\left(k_{Q_M}\right)^{n_M}$ sposobów, a dodatkowo muszą wziąć pod uwagę, że usługa zostanie wyświadczona przez rzetelnego agenta, stąd kolejne $n_B - 1$ przypadków. Wszystko to sprawia, że dokonanie analizy możliwych przypadków jest niemożliwe w niektórych rodzajach systemów TRM (w systemach, w których agenty mogą wydawać bardzo wiele rekomendacji np. dotyczących wszystkich innych agentów). Dlatego można pokusić się o uproszczenie – wydawanie takich samych rekomendacji przez wszystkie agenty złośliwe – jest to kolejna heurystyka – nie analizujemy wszystkich możliwych przypadków wydawanych rekomendacji przez każdego złośliwego agenta z osobna ale wszystkie razem w taki sam sposób. Wtedy liczba przypadków będzie istotnie mniejsza i wyniesie: $k_{R_{A_B}} * k_{R_{A_M}} * k_{Q_M}$ przypadków (zakładamy, że wtedy agenty złośliwe nie rozważają osobno przypadku, w którym to agent rzetelny będzie świadczył usługę, ale wezmą pod uwagę w danym przypadku prawdopodobieństwo tego, że to agent

rzetelny zostanie wybrany jako usługodawca oraz prawdopodobieństwo wyboru agenta złośliwego jako usługodawcę).

5.4.5. Analiza możliwych zachowań atakujących jedynie odnośnie świadczenia usługi

Ze względu na to, że analizowanie rekomendacji złośliwych agentów może generować bardzo wiele niezależnych przypadków, to celowym może być pominięcie tego aspektu i przyjęcie, że agenty złośliwe dostarczają rekomendacje o stałej wartości w stosunku do każdego agenta złośliwego i każdego agenta rzetelnego (także zaniżonej lub zawyżonej) podczas działania środowiska. Wtedy istnieje konieczność jedynie przeanalizowania decyzji związanej z jakością świadczonej usługi. Co więcej, analizę tę można przeprowadzić w przypadku gdy inny agent już zdecyduje się na skorzystanie z usługi dostarczanej przez złośliwego agenta, co znacząco ogranicza liczbę przypadków do przeanalizowania. Wtedy, w ramach metody MEAEM rozważane są jedynie wartości jakości usług, które mogą zostać wyświadczone dla agentów: $q_i \in Q_M \subset Q$.

5.4.6. Funkcja zysku atakujących

Przyjmijmy, ze celem jest minimalizacja efektywności środowiska w czasie całego działania środowiska. Aby było możliwe dokonanie wyboru parametrów decyzji atakujących metodą MEAEM, które umożliwiają w jak największym stopniu realizację tego celu, konieczne jest wprowadzenie definicji funkcji zysku atakujących, a następnie określenie wyrażenia umożliwiającego obliczenie jej wartości.

Definicja 5.4.6. ***Funkcją zysku atakujących*** *jest funkcja* $f_g: Q_M \times R_{A_M} \times R_{A_B} \times M \to \mathbb{R}$, $f_g\left(q_i, r_i^{A_M}, r_i^{A_B}, m_i\right) = g_i$ *przy czym* $q_i \in Q_M$ *– jakość usługi świadczonej przez złośliwego agenta,* $r_i^{A_M} \in R_{A_M}$ *– wartość rekomendacji agentów złośliwych o agentach złośliwych,* $r_i^{A_B} \in R_{A_B}$ *– wartość rekomendacji agentów złośliwych o agentach rzetelnych,* $m_i \in M$ *– czas interakcji o numerze i,* $g_i \in \mathbb{R}$ *– wartość zysku atakujących.*

Przyjmijmy następujące oznaczenia, potrzebne do stworzenia propozycji obliczania funkcji zysku atakujących[60]:

- $\%_{A_B}^{m_l}$ – prawdopodobieństwo wyboru agenta rzetelnego jako usługodawcę w interakcji m_l;
- $\%_{A_M}^{m_l}$ – prawdopodobieństwo wyboru agenta złośliwego jako usługodawcę w interakcji m_l;
- $\Delta t_{A_B \to A_B}^{c_k;m_l} = t_{A_B \to A_B}^{c_k;m_l} - t_{A_B \to A_B}^{c_k;m_{l-1}}$ – zmiana globalnego średniego zaufania agentów rzetelnych do agentów rzetelnych w kontekście c_k po interakcji m_l, przy czym:
  - $\Delta t_{A_B \to A_B/A_B}^{c_k;m_l}$– zmiana globalnego średniego zaufania agentów rzetelnych do agentów rzetelnych w kontekście c_k po interakcji m_l, w przypadku gdy usługodawcą w tej interakcji był agent rzetelny,
  - $\Delta t_{A_B \to A_B/A_M}^{c_k;m_l}$– zmiana globalnego średniego zaufania agentów rzetelnych do agentów rzetelnych w kontekście c_k po interakcji m_l, w przypadku gdy usługodawcą w tej interakcji był agent złośliwy;
- $\Delta t_{A_B \to A_M}^{c_k;m_l} = t_{A_B \to A_M}^{c_k;m_l} - t_{A_B \to A_M}^{c_k;m_{l-1}}$ – zmiana globalnego średniego zaufania agentów rzetelnych do agentów złośliwych w kontekście c_k po interakcji m_l, przy czym:
  - $\Delta t_{A_B \to A_M/A_B}^{c_k;m_l}$– zmiana globalnego średniego zaufania agentów rzetelnych do agentów złośliwych w kontekście c_k po interakcji m_l w przypadku gdy usługodawcą w tej interakcji był agent rzetelny,
  - $\Delta t_{A_B \to A_M/A_M}^{c_k;m_l}$– zmiana globalnego średniego zaufania agentów rzetelnych do agentów złośliwych w kontekście c_k po interakcji m_l w przypadku gdy usługodawcą w tej interakcji był agent złośliwy;
- $\Delta E^{(n):m_l} = E^{(n):m_l} - E^{(n):m_{l-1}}$ – zmiana efektywności chwilowej n środowiska po interakcji m_l, przy czym:
  - $\Delta E_{/A_B}^{(n):m_l}$ – zmiana efektywności chwilowej n środowiska po interakcji m_l w przypadku gdy usługodawcą w tej interakcji był agent rzetelny;
  - $\Delta E_{/A_M}^{(n):m_l}$ – zmiana efektywności chwilowej n środowiska po interakcji m_l w przypadku gdy usługodawcą w tej interakcji był agent złośliwy.
- γ_E – współczynnik istotności efektywności środowiska, $\gamma_E \in \mathbb{R}_+$;

[60] Dalsze rozważania zostaną prowadzone w oparciu o pojęcie zaufania, w przypadku gdy dany system TRM korzysta z reputacji zamiast zaufania, poszczególne wartości i wyrażenia dotyczące zaufania należy zastąpić odpowiednimi wyrażeniami reputacji.

- $\gamma_{t^{c_k}}$ – współczynnik istotności zaufania w kontekście c_k, przy czym: $\gamma_{t^{c_k}} \in \mathbb{R}_+$.

Wartości powyższych parametrów, z wyjątkiem współczynników istotności, są wyznaczane dla każdej interakcji i każdego zestawu decyzji. Wartości współczynników istotności są wyznaczane raz dla danego użycia metody MEAEM.

Jak zauważono wcześniej, korzystny dla atakujących jest spadek efektywności środowiska, spadek zaufania agentów rzetelnych do agentów rzetelnych (w różnych kontekstach), wzrost zaufania agentów rzetelnych do złośliwych (w różnych kontekstach) oraz wzrost prawdopodobieństwa wyboru agenta złośliwego jako usługodawcy (umożliwi to bowiem wykonywanie kolejnych ataków, ale prawdopodobieństwo to powinno w danym systemie być związane z wartością zaufania do tego agenta). Na tej podstawie można skonstruować następujące wyrażenie na funkcję zysku atakujących, podczas podejmowania decyzji dotyczącej dostarczania rekomendacji i jakości świadczonej usługi:

Wzór 5.1. ***Funkcja zysku atakujących****, podczas podejmowania decyzji dotyczącej dostarczania rekomendacji i jakości świadczonej usługi:*

$$f_g{}^{r+a} = \%_{A_M}^{m_l}\left(-\gamma_E\,\Delta E_{/A_M}^{(n):m_l} - \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_B/A_M}^{c_k;m_l} + \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_M/A_M}^{c_k;m_l}\right) + \%_{A_B}^{m_l}\left(-\gamma_E\,\Delta E_{/A_B}^{(n):m_l} - \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_B/A_B}^{c_k;m_l} + \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_M/A_B}^{c_k;m_l}\right)$$

Można także skonstruować następujące wyrażenie na funkcję zysku atakujących podczas podejmowania decyzji dotyczącej jedynie jakości świadczonej usługi:

Wzór 5.2. ***Funkcja zysku atakujących****, podczas podejmowania decyzji dotyczącej jakości świadczonej usługi:*

$$f_g{}^{a} = -\gamma_E\,\Delta E_{/A_M}^{(n):m_l} - \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_B/A_M}^{c_k;m_l} + \sum_k \gamma_{t^{c_k}} * \Delta t_{A_B\to A_M/A_M}^{c_k;m_l}$$

Warto zauważyć, że po wybraniu zestawu decyzji przed wydaniem rekomendacji, należy dokonać weryfikacji wybranej wcześniej jakości usługi z uwagi na fakt, że dla konkretnego złośliwego agenta, który będzie świadczył usługę, ze względu na możliwą jego odmienną charakterystykę (np. zaufanie usługobiorcy do niego) optymalna decyzja dla niego dotycząca wyboru jakości usługi może być inna niż w przypadku średnich wartości. W danej interakcji agenty złośliwe powinny stosować takie zachowanie, które umożliwi osiągnięcie maksymalnej wartości funkcji zysku atakujących.

W przypadku gdy liczba wydawanych rekomendacji nie jest zbyt duża (tzn. np. rekomendacja nie jest wydawana przez każdego agenta i nie na temat każdego innego agenta) oraz liczba złośliwych agentów jest znacząco mniejsza od liczby agentów rzetelnych, warunek ten można dodatkowo rozbudować i analizować oddzielnie rekomendację każdego ze złośliwych agentów. Wtedy we wzorze 5.1. należy zamienić prawdopodobieństwo wyboru agenta złośliwego jako usługodawcę w interakcji m_l, na wektor prawdopodobieństw wyboru poszczególnych agentów, a zmiany globalnego średniego zaufania w przypadku gdy usługodawcą był agent złośliwy lub rzetelny, na wektory zaufania do każdego agenta (a nie do całej grupy agentów). Wtedy wybór złośliwych agentów może okazać się lepszy. Jednak wydaje się, że w praktyce realizacja obliczeń byłaby możliwa jedynie dla niewielkich środowisk i dodatkowo jedynie w przypadku spełnienia powyższych warunków (ograniczenia liczby rekomendacji), wobec tego praktyczne zastosowania takiego sposobu byłyby niewielkie.

O ile to, że na wartość funkcji zysku atakujących powinna mieć wpływ zmiana efektywności chwilowej w związku z decyzjami rozważanymi w danej interakcji, jak i wartości zaufania do agentów rzetelnych i złośliwych ze strony agentów rzetelnych oraz prawdopodobieństwo tego, że w danej interakcji usługę będzie świadczył agent rzetelny lub złośliwy, wydaje się być dość naturalne i logicznie uzasadnione, to problemem jest dobór współczynników γ_E oraz $\gamma_{t^{c_k}}$ (dla każdego kontekstu zaufania c_k). Zgodnie z propozycją autora niniejszej pracy współczynniki te należy dobrać w taki sposób, aby zakres zmienności poszczególnych czynników powyższych wyrażeń był zbliżony. Oceny zakresu zmienności można dokonać dla kilku pierwszych interakcji[61]. Dodatkowo należy wziąć pod uwagę fakt, czy wszystkie konteksty zaufania są od siebie w pełni niezależne, a w przypadku gdy tak nie jest należy dokonać odpowiedniego zmniejszenia wartości tych czynników.

[61] Autor pomija stosowny przykład w tym miejscu, gdyż dobór tych współczynników zostaje dokonany w trakcie badań systemu RefTRM w punkcie 6.3.8.

6. BADANIE SYSTEMU TRM W OPARCIU O METODYKĘ OCENY WIARYGODNOŚCI

Niniejszy rozdział przedstawia narzędzie stworzone do oceny wiarygodności systemów TRM, a także prezentuje wyniki badań przeprowadzonych (w oparciu o metodykę oceny wiarygodności) dla wybranego, przykładowego systemu TRM.

6.1. NARZĘDZIE DO OCENY WIARYGODNOŚCI SYSTEMÓW TRM

Przyjęto, że tworzone narzędzie umożliwiające ocenę wiarygodności systemów zarządzania zaufaniem powinno spełniać następuje wymagania:

- możliwość łatwej implementacji nowych systemów TRM,
- możliwość łatwej implementacji nowych ataków,
- definicja zestawu testów (ataków w stosunku do różnych środowisk w jakich działa system TRM), umożliwiającego porównywanie efektywności różnych systemów,
- implementacja zdefiniowanych miar efektywności i badań systemów TRM w oparciu o metodykę oceny wiarygodności.

W celu spełniania powyższych wymagań zostało wykonane narzędzie TRM-RET (Trust and Reputation Management – Reliability Evaluation Testbed), opracowane na podstawie wcześniej stworzonego przez autora niniejszej pracy narzędzia TRM-EAT, opisanego w artykule [109] oraz jego wcześniejszej wersji opisanej przez autora rozprawy w artykule [93]. Narzędzie TRM-RET zostało zaimplementowane w języku Python, a konfiguracja badań odbywa się za pomocą pliku w formacie JSON. Zestawienie bibliotek i zasobów wymaganych do uruchomienia narzędzia zostały wyszczególnione w załączniku 5.

W kolejnych punktach opisano architekturę i ogólny sposób funkcjonowania narzędzia oraz sposoby implementacji systemów TRM i ataków na nie, a także sposób przeprowadzania badań z użyciem tego narzędzia.

6.1.1. Architektura

Do najistotniejszych klas obiektów zaimplementowanych w narzędziu należą:

- klasa *Agent*, która reprezentuje agenta w środowisku,
- klasa *Environment*, która reprezentuje środowisko,
- klasa *Trust*, która reprezentuje zaufanie danego agenta do innego określonego agenta lub reputację określonego agenta,
- klasa *RMSS*, która reprezentuje podsystem zarządzania rekomendacjami.

Zgodnie z uwagą 4.2.8., w środowisku w którym działa system TRM, obsługa żądania świadczenia usługi składa się z następujących kroków:

- pozyskanie rekomendacji;
- ocena zaufania lub reputacji;
- wybór usługodawcy;
- interakcja;
- wystąpienie zakłócenia;
- obserwacja (jeżeli jest możliwa w danym środowisku);
- ocena rekomendacji;
- ocena lub aktualizacja zaufania lub reputacji.

Mechanizm działania narzędzia odzwierciedla przedstawione powyżej i we wcześniejszych rozdziałach pracy, etapy obsługi żądania, i co za tym idzie, etapy działania systemu TRM i środowiska. Podstawowym obiektem w symulatorze jest *agent* (obiekt klasy *Agent*). Każdy *agent* implementuje następujące metody, opisane w artykule [93] autora rozprawy:

- *„get_recommendations() – żąda rekomendacji od innych agentów na temat określonego agenta (lub agentów),*
- *calculate_trust() – oblicza zaufanie do poszczególnych agentów na bazie otrzymanych rekomendacji oraz historii wcześniejszych interakcji lub wartości zaufania do innych agentów,*
- *choose_provider() – dokonuje wyboru agenta, który będzie świadczył usługę. Sposób wyboru agenta jest uzależniony od systemu zarządzania zaufaniem, ale jest oparty na wartościach zaufania obliczonych w poprzedniej metodzie,*
- *get_service() – żąda usługi od wcześniej wybranego agenta,*
- *eval() – dokonuje oceny jakości otrzymanej usługi oraz uaktualnia wartości zaufania do innych agentów, a także uaktualnia historię rekomendacji."*

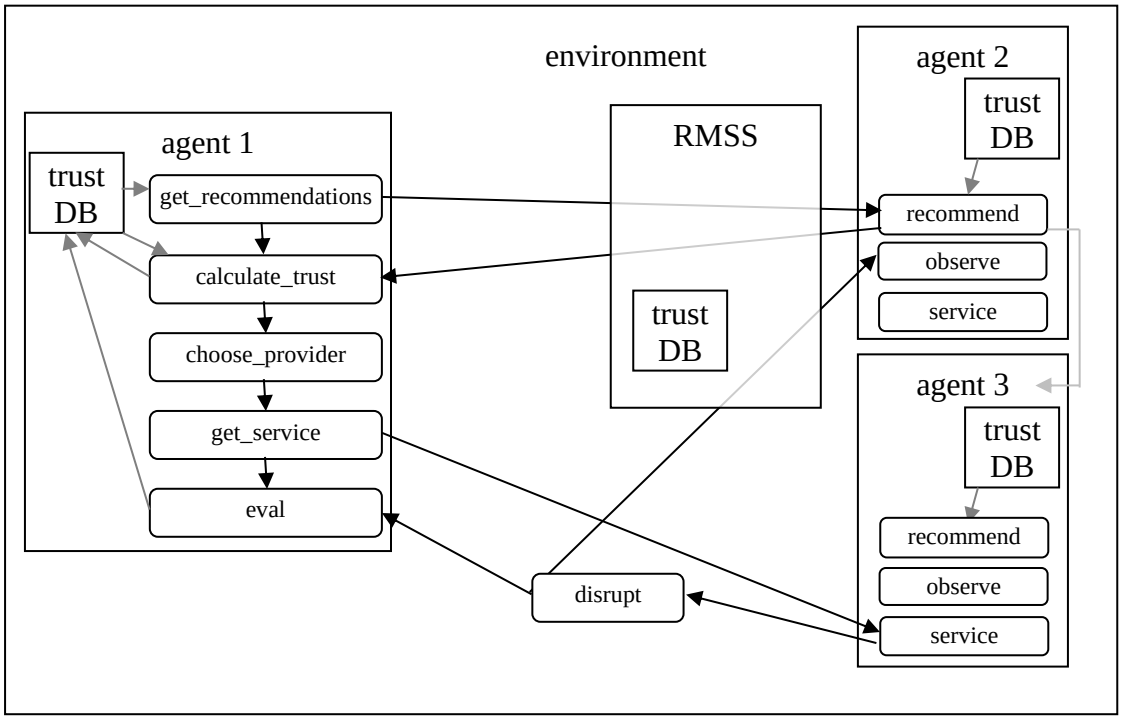
Dodatkowo każdy agent implementuje też następujące metody, wymienione w artykule [93]:

- *„recommend(agent, service) – zwraca rekomendację dotyczącą zaufania do agenta agent przy świadczeniu usługi określonej jako service,*
- *service(service) – zwraca wynik (jakość) działania usługi określonej jako service,*
- *observe(agent, service) – umożliwia rejestrację obserwacji (wyników interakcji pomiędzy innymi agentami), o ile są one możliwe w danym środowisku."*

Klasa *Environment* posiada metodę *disturb()* – modelującą występowanie zakłóceń (wystąpienie zakłócenia powoduje zmianę jakości usługi dostarczonej do usługobiorcy w stosunku do usługi świadczonej przez usługodawcę).

Istnieje także obiekt klasy *RMSS* – służący do przekazywania lub obsługi żądań związanych z wydawaniem rekomendacji oraz ewentualnie związanych z rejestrowaniem obserwacji. Każdy *agent* oraz obiekt klasy *RMSS* posiada obiekt *trust DB*, w którym są przechowywane informacje służące do obliczenia wartości zaufania i reputacji dla poszczególnych agentów, mogą one obejmować m.in. wyniki własnych interakcji, obserwacje, otrzymane rekomendacje itd.

Logika działania narzędzia została opisana poniżej i przedstawiona schematycznie na rysunku 9.



Rysunek 9 Logika działania narzędzia TRM-RET

Działanie narzędzia zostało opisane we wcześniejszym artykule [93] autora niniejszej rozprawy: *„załóżmy, że agent 1 chce skorzystać z pewnej usługi udostępnionej przez agenta 3. W tym celu agent 1 chce poznać rekomendacje innych agentów na temat agenta 3. Za pomocą metody get_recommendations() dokonuje on*[62] *selekcji agentów, którzy dostarczą taką rekomendację. Załóżmy, że agentem dostarczającym rekomendację będzie agent 2. Wtedy agent 1 wywołuje metodę recommend() udostępnioną przez agenta, który ma dostarczyć*

[62] samodzielnie lub z użyciem podsystemu RMSS

rekomendację. Warto podkreślić, że wybranie zbioru agentów od których dany agent chce uzyskać rekomendacje, a także określenie jak często i w jakich sytuacjach rekomendacje będą pozyskiwane, zależy od konkretnego systemu zarządzania zaufaniem, a dokładniej od implementacji metody get_recommendations(). W odpowiedzi agent 2 zwraca rekomendację na temat agenta 3.[63] *Na bazie uzyskanych rekomendacji, a także przechowywanej bazy zaufania (trust DB) obejmującej np. historię poprzednich interakcji z innymi agentami, agent 1 dokonuje obliczenia zaufania do potencjalnych usługodawców w ramach metody calculate_trust(), której implementacja zależy od konkretnego systemu zarządzania zaufaniem. Po wykonaniu kalkulacji mogą zostać uaktualnione wpisy w bazie zaufania. Następnie, agent 1 wybiera agenta od którego zażąda wykonania usługi poprzez wywołanie metody choose_provider(). Sposób wyboru jest uzależniony od konkretnego systemu TRM – może to być np. agent darzony najwyższym zaufaniem, albo wylosowany agent spośród tych, do których zaufanie przekracza ustaloną wartość. Następnie, agent 1 żąda usługi (wywołuje własną metodę get_service()) poprzez wywołanie metody service() udostępnionej przez wybranego agenta (na rysunku jest to agent 3). Agent świadczący usługę zwraca jej wynik. W środowisku mogą wystąpić zakłócenia, które wpłyną na ten wynik (służy do tego metoda disrupt(), której implementacja zależy od konkretnego środowiska). W przypadku gdy możliwe są obserwacje, zostaje wywołana metoda observe() u agentów, którzy mogli dokonać obserwacji. Następnie, agent 1 wykonuje metodę eval(), która ma za zadanie odpowiednio uaktualnić zaufanie do agenta świadczącego usługę, ale także do innych agentów (np. tych, które wydały rekomendacje na temat agenta 3). W kolejnych krokach symulacji jest wybierany kolejny agent, który chce skorzystać z usługi oraz ta usługa, a następnie poszczególne działania związane z obsługą takiego żądania przebiegają analogiczne jak przedstawiono powyżej."*

6.1.1. Sposób implementacji środowiska

W celu stworzenia środowiska, konieczne jest utworzenie klasy dziedziczącej po klasie *environment*, której definicja będzie zawierała odpowiednią liczbę agentów, połączeń pomiędzy nimi, a także świadczonych usług oraz charakterystyki poszczególnych agentów. Istotna jest także implementacja metody *disturb()*, symulującej występowanie zakłóceń oraz metody *observe()* umożliwiającej dokonywanie obserwacji przez agentów.

[63] Zarówno żądanie rekomendacji jak i sama rekomendacja mogą zostać dokonane także za pośrednictwem podsystemu RMSS, o ile użycie takiego komponentu jest wymagane przez dany system TRM (w przypadku większości systemów TRM rola podsystemu RMSS sprowadza się jedynie do przekazywania komunikatów, co zostało szerzej omówione w punkcie 4.2.3.6.).

6.1.2. Sposób implementacji systemów TRM

W artykule [93] autor rozprawy, zaznaczył, że w celu implementacji systemu TRM: *„należy utworzyć nową klasę agentów, dziedziczącą po klasie Agent i zaimplementować następujące metody:*

- *get_recommendations(),*
- *calculate_trust(),*
- *choose_provider(),*
- *get_service(),*
- *eval().”*

Możliwa jest także implementacja klasy dziedziczącej po *RMSS*. Obiekt tej klasy będzie wtedy odpowiedzialny za zarządzanie rekomendacjami agentów według danego systemu TRM.

6.1.3. Sposób implementacji ataków

W artykule [93] autor niniejszej rozprawy podkreślił, że: *„stworzenie nowego ataku wymaga implementacji dwóch metod tj. recommend() oraz service() w klasie Agent. Występuje więc niezależność implementacji systemu TRM i ataków na te systemy (w przypadku większości ataków, które są specyficzne dla danego systemu TRM). Dzięki temu, że wybrano taki sposób implementacji nowych ataków, to w przypadku chęci uwzględnienia w narzędziu nowego systemu zarządzania zaufaniem i reputacją, nie ma potrzeby definiowania sposobu działania złośliwych agentów dla każdego z ataków z osobna (istnieje rozdział pomiędzy implementacją systemu zarządzania zaufaniem, a implementacją ataków na ten system).”*

6.1.4. Zaimplementowane ataki

W narzędziu zaimplementowano najistotniejsze (m.in. takie, które mogą być zastosowane przeciwko szerokiej klasie systemów TRM) ataki spośród tych, które zostały wymienione w literaturze i opisane w załączniku 4, tj.:

- oczernianie (ozn. **B** – ang. Bad-mouthing);
- wychwalanie (ozn. **F** – ang. False-praise);
- atak stały (ozn. **C** – ang. Constant);
- oscylacja zachowania (ozn. **O** – ang. On-off);
- niespójne zachowanie (ozn. **N** – ang. coNflicting bahaviour)
- wyrocznia (ozn. **W**).

Możliwe jest wykonywanie powyższych ataków przez pojedynczych agentów, ale zostały stworzone także wersje tych ataków wykorzystujące kooperację, czyli współpracę wielu agentów (w przypadkach, w których oryginalna wersja ataku nie zakładałaby kooperacji, czy innej formy współpracy złośliwych agentów). Zaimplementowane zostały także kombinacje powyższych ataków, takie jak.

- atak oczerniania i wychwalania (ozn. **BF**);
- atak oczerniania, wychwalania i stały z kooperacją (ozn. **BFCc**);
- atak oczerniania, wychwalania i oscylacji zachowania z kooperacją (ozn. **BFOc**)[64].

Został także zaimplementowany atak o nazwie MEAEM (ozn. **M**), który służy do przeprowadzenia badań zgodnie z opisaną w podrozdziale 5.4. metodą. Możliwe jest także wykorzystanie w ramach wszystkich powyższych ataków, ataku kreacji wielu tożsamości (ozn. **S** – ang. sybil) oraz ataku kreacji nowej tożsamości (ozn. **H** – ang. whitewashing), ale możliwość przeprowadzenia tych ataków musi być jasno wskazana w momencie konfiguracji badania, gdyż zależy ona przede wszystkim od środowiska, w którym działa system TRM, a w niewielkim stopniu od samego systemu TRM.

Niektóre ataki (C, B, F, BF, BFC) nie wymagają dodatkowych parametrów. Inne (O, N, W, BFO) przyjmują dodatkowe parametry. Zestawienie ataków wraz z parametrami domyślnymi zaimplementowanymi w narzędziu przedstawiono w tabeli 2. Jeżeli nie zaznaczono inaczej, w dalszych badaniach są używane domyślne parametry ataków.

Tabela 2 Parametry ataków

Atak	**Parametr (nazwa w TRM-RET)**	**Zakres**	**Wartość domyślna**	**Wyjaśnienie**
O W BFO	współczynnik (ratio)	(0,1)	0.4	Wszystkie te ataki używają takich samych parametrów. Interwał określa długość cyklu (liczbę usług wyświadczonych w cyklu), a współczynnik, jaka część z tych usług będzie miała zaniżoną jakość. Współczynnik 0.4 i interwał 5 oznaczają, że agent wyświadczy 3 usługi z najwyższą możliwą jakością, kolejne 2 z najniższą jakością, po czym znowu 3 usługi z najwyższą itd.
	interwał (interval)	ℕ	5	

[64] Ataki FBCc i FBOc są uogólnieniem ataków: złośliwy kolektyw, złośliwy kolektyw z kamuflażem, złośliwi szpiedzy, obniżenie reputacji dobrych agentów i częściowo złośliwy kolektyw, które zostały opisane w artykule [74] oraz przytoczone w załączniku 4.

N	Lista agentów wobec których są świadczone usługi o najniższej jakości (list_of_agents_off')	Lista zawierająca $1 \ldots n_B$ identyfikatorów rzetelnych agentów	[1,2,3,4,5]	Lista identyfikatorów agentów wobec których są świadczone usługi o najniższej jakości; pozostałym agentom rzetelnym, agenty złośliwe świadczą usługi o wysokiej jakości.

6.1.5. Konfiguracja badań

Badania są przeprowadzane w oparciu o zdefiniowany plik konfiguracyjny (w formacie json). Przykładowy plik konfiguracyjny zaprezentowano na listingu 1.

```
{
    "simulationName": "Generic simulation",
    "nrSimulations": 3,
    "environment": {
        "services": [
            {
                "name": "S1",
                "description": "default service S1"
            },
            {
                "name": "S2",
                "description": "additional service S2"
            }
        ],
        "nrAgents": 20,
        "agents": [
            {
                "agents": "all",
                "type": "agentTypeDefault",
                "servicesOffered": [
                    "S1",
                    "S2"
                ],
                "QoS": "max"
            }
        ],
        "characteristic": "online",
        "links": [
        ]
    },
    "attacks": {
        "nrAgentsMalicious": 10,
        "cooperation": true,
        "attack types": [
            {
                "type": "BFC",
                "number": 10
            }
        ]
    },
    "interactions": {
        "nrInteractions": 1000,
        "requestsType": "random",
    },
    "nrTRMs": 1,
    "TRMs": [
        {
            "name": "ReferenceTrustModel",
            "parameters": {
                "threshold": 0.5
            }
        }
    ]
}
```

Listing 1 Przykład pliku konfiguracyjnego

Konfiguracja jest dokonywana w kilku obszarach [93]:

1. Podstawowe parametry dotyczące badania, takie jak liczba dokonywanych symulacji (dla jednego zestawu parametrów), nazwa badania, itd.
2. Parametry środowiska, takie jak:
   - określenie liczby agentów w środowisku;
   - definiowanie topologii środowiska (sieci), w której ma działać system TRM, istnieje możliwość stworzenia sieci ze strukturą:
     - losową lub full-mesh,
     - hierarchiczną, w oparciu o określone parametry, np. liczbę stopni hierarchii,
     - zdefiniowaną w pliku;
3. Określenie parametrów przebiegu symulacji (np. czas trwania symulacji, liczbę interakcji pomiędzy agentami do zakończenia symulacji, czy sposób generowania żądań).
4. Określenie systemu zarządzania zaufaniem i jego parametrów.
5. Określenie ataków i ich parametrów, którym będzie poddany dany system zarządzania zaufaniem.

6.1.6. Prezentacja wyników badań

Wyniki badań zawarte są w postaci wykresów prezentujących określone miary oraz w postaci pliku json, którego przykładowy fragment został zaprezentowany na listingu 2. W wynikach znajdują się wartości miar zdefiniowanych w rozdziale 5, wyznaczone po każdej interakcji, a także wybrane parametry agentów (np. możliwe jest obserwowanie wartości zaufania pomiędzy dowolna parą agentów).

```
{
    "statistics": {
        "nrInteraction": [0,1,2,3,4],
        "outcome": [1.0,0.0,1.0,0.0,1.0],
        "agentProvider": [6,13,9,3,18],
        "agentClient": [1,12,4,17,1],
        "trustBenToBen": [0.5,0.6,0.5,0.45,0.41]
}
```

Listing 2 Fragment przykładowego pliku wynikowego

6.2. OPIS BADANEGO SYSTEMU I ŚRODOWISKA

W niniejszym podrozdziale przedstawiono najważniejsze informacje dotyczące środowiska i systemu TRM poddanego badaniom, jak i ogólną charakterystykę wykonywanych badań.

6.2.1. Środowisko

Badane środowisko składa się z 20 agentów ($n = 20$ – liczba agentów w systemie): $A = \{a_1, a_2, \dots, a_{20}\}$, w którym świadczona jest jedna usługa: $U = \{u_1\}$. Wszystkie agenty mogą być zarówno usługodawcami, jak i usługobiorcami, czyli: $A_{P:u_l} = \{a_1, \dots, a_{20}\}$, $\forall_{k\epsilon\mathbb{N},1\leq k\leq 20}\ U_{a_k} = \{u_1\}$. Każdy agent może żądać świadczenia usługi od każdego z agentów świadczących daną usługę[65] (mamy do czynienia z połączeniami każdego agenta z każdym agentem-usługodawcą na poziomie warstwy usługowej), przy czym nie może on żądać świadczenia usługi od samego siebie. Usługi w środowisku mogą być świadczone z jakością $Q =< 0{,}1 >$, gdzie $q = 0$ odpowiada brakowi usługi, a $q = 1$ wyświadczeniu usługi o najwyższej jakości. Każdy z usługodawców jest w stanie świadczyć usługę z maksymalną jakością $q = 1$, czyli: $\forall_{k\epsilon\mathbb{N},1\leq k\leq 5}\ q_{a_k}^{u_1} = 1$. W przypadku gdy żaden z agentów nie stosuje ataku, każdy z agentów świadczy usługi z maksymalną możliwą jakością, $q = 1$, czyli: $\forall_{l,k}\ f_{intE}(e_l, a_k) = 1$. W środowisku nie występują zakłócenia, czyli: $\forall_l\ f_{dis}(z_l, m_l) = q^l$, a więc: $\forall_l\ q^l = o^l$.

6.2.2. System RefTRM

Badaniom zostanie poddany system RefTRM, opisany jako przykład w punkcie 4.2.9. Z tego względu opis systemu nie został powielony, a jedynie w tabeli 3 zostały zaprezentowane wartości parametrów jakie używa ten system.

Tabela 3 Domyślne wartości parametrów systemu RefTRM

Oznaczenie	Nazwa lub opis parametru	Wartość przyjęta w badaniach (jeżeli dalej nie zaznaczono inaczej)
$t_{init}^{c_1}$	początkowe zaufanie akcyjne	0.5
$t_{init}^{c_2}$	początkowe zaufanie rekomendacyjne	0.5
α	waga zaufania akcyjnego	0.7
h	próg minimalnego zaufania	0.5

[65] W ramach tego środowiska, mamy do czynienia z sytuacją, że wszyscy agenci zarówno świadczą, jak i żądają świadczenia tej samej usługi. Taka sytuacja, rzecz jasna, nie wystąpi w rzeczywistych środowiskach, niemniej jednak, ułatwi interpretację wyników badań, a z drugiej strony nie wpłynie negatywnie na rzetelność tych wyników. W łatwy sposób możliwe jest dokonanie jasnego podziału na usługodawców i usługobiorców, ale ze względu na to że w środowisku występuje niewielka liczba agentów, autor zdecydował się na odstąpienie od tego.

β^S	zmniejszenie prawdopodobieństwa wyboru na dostawcę usługi	0.1
α^R	zwiększenie zaufania rekomendacyjnego za prawidłową rekomendację	0.1
β^R	zmniejszenie zaufania rekomendacyjnego za nieprawidłową rekomendację	0.25
h^R	próg prawidłowości rekomendacji	0.6
α^A	zwiększenie zaufania akcyjnego za satysfakcjonującą jakość usługi	0.2
β^A	zmniejszenie zaufania akcyjnego za niesatysfakcjonującą jakość usługi	0.4
h^A	próg satysfakcjonującej usługi	0.5

6.2.3. Ogólna charakterystyka badań

O ile w przypadku konkretnego badania nie zaznaczono inaczej, badania cechują się następującą charakterystyką:

- w środowisku jest 20 agentów (w przypadku ataków 10 z nich to agenty złośliwe);
- długość symulacji dla każdego badania wynosi: 1000 interakcji;
- charakterystyka żądań jest ustalona dla wszystkich badań (wylosowana jednokrotnie);
- dla każdego badania wykonano przynajmniej 10 przebiegów badania (dla każdego parametru zaprezentowano wartość średnią z przebiegów symulacji oraz wartość minimalną i maksymalną).

6.3. OCENA WIARYGODNOŚCI SYSTEMU REFTRM

Niniejszy podrozdział prezentuje wyniki oceny wiarygodności systemu RefTRM (zdefiniowanegp w punkcie 6.2.2) w środowisku określonym w punkcie 6.2.1. Ocena wiarygodności została wykonana w oparciu o metodykę, miary i rodzaje badań, zdefiniowane w rozdziale 5, z wykorzystaniem narzędzia TRM-RET, opisanego w podrozdziale 6.1.

6.3.1. Wstępna ocena podatności na ataki

Przyjęto, że w środowisku nie jest możliwe wykonanie ataku kreacji wielu tożsamości lub kreacji nowej tożsamości, gdyż w środowisku mogą funkcjonować jedynie agenty istniejące od początku jego działania. Wobec tego, ataki wykorzystujące kreację wielu albo nowej tożsamości, nie będą badane.

W odniesieniu do tego systemu TRM i środowiska, zastosowanie będą miały miary wiarygodności wyszczególnione w tabeli 4. Ze względu na to, że określając środowisko

przyjęto, że każdy z agentów jest w stanie świadczyć usługi z maksymalną jakością, więc zarówno idealna efektywność E^{ideal}, jak i zaawansowana efektywność E^{adv}, będą w każdym momencie równe efektywności środowiska E. Dodatkowo, w środowisku nie występują zakłócenia, wobec czego efektywność środowiska bez zakłóceń E_q będzie tożsama z efektywnością środowiska E. Wobec tego miary te nie będą prezentowane i zostały wyszarzone w tabeli 4. Miarę efektywności chwilowej n można zaprezentować dla różnej wartości parametru n, ale w wynikach badań zaprezentowano[66] wartość tej miary dla $n = 100$, o ile nie zaznaczono inaczej. Miary zysku efektywności G i zysku absolutnego efektywności G_A zostaną wyznaczone po przeprowadzeniu wszystkich badań.

Podczas badań środowiska bez działającego systemu TRM prezentowane będą jedynie miary efektywności środowiska oraz efektywności chwilowej n. Wszystkie miary uwzględniające wartości zaufania nie mają zastosowania z uwagi na to, że bez użycia systemu TRM nie są określane wartości zaufania.

Tabela 4 Miary wiarygodności zastosowane w badaniach systemu RefTRM

E_q	efektywność środowiska bez zakłóceń
E	efektywność środowiska
$E^{(n)}$	efektywność chwilowa n
E^{ideal}	idealna efektywność
E^{adv}	zaawansowana efektywność
$l_I^{A_B,A_B}$	liczba interakcji agentów rzetelnych z agentami rzetelnymi
$l_I^{A_B,A_M}$	liczba interakcji agentów rzetelnych z agentami złośliwymi
$t_{A_B\to A}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do wszystkich agentów[67]
$t_{A_B\to A}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do wszystkich agentów
$t_{A_B\to A}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do wszystkich agentów
$t_{A_B\to A_B}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do agentów rzetelnych
$t_{A_B\to A_B}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów rzetelnych
$t_{A_B\to A_B}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do agentów rzetelnych
$t_{A_B\to A_M}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do agentów złośliwych

[66] Wybór wartości $n = 100$ jest nieprzypadkowy, w kontekście warunków panujących w środowisku oraz liczby interakcji jakie w nim zachodzą. Z jednej strony ideą miary efektywności chwilowej jest wyeliminowanie długoterminowego wpływu początkowych warunków panujących w środowisku (co skłania ku małym wartościom n), a z drugiej strony parametr ten nie może być zbyt mały, gdyż istotne jest zmniejszenie i uśrednienie wpływu losowych właściwości środowiska oraz stworzenie miary, która nie będzie zanadto zmienna. Przyjęta wartość jest więc pewnego rodzaju konsensusem.

[67] W tej i kolejnych miarach użyto, za przykładem 5.1.3.1., pojęcia zaufania akcyjnego, rekomendacyjnego i całkowitego, odpowiednio dla zaufania w kontekście c_1, c_2 i c_3

$t_{A_B \to A_M}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów złośliwych
$t_{A_B \to A_M}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do agentów złośliwych
G	zysk efektywności
G_A	zysk absolutny efektywności

W przypadku środowiska z działającym systemem RefTRM, symulowane będą wszystkie ataki, jakie zostały zaimplementowane w narzędziu, tj.

- atak stały (ozn. **C** – ang. Constant);
- oczernianie i wychwalanie[68] (ozn. **BF** – ang. Bad-mouthing, False-praise);
- oscylacja zachowania (ozn. **O** – ang. On-off);
- niespójne zachowanie (ozn. **N** – ang. coNflicting bahaviour);
- wyrocznia (ozn. **W**);
- atak oczerniania, wychwalania i stały z kooperacją (ozn. **BFCc**);
- atak oczerniania, wychwalania i oscylacji zachowania z kooperacją (ozn. **BFOc**).

W przypadku środowiska bez systemu TRM możliwe jest przeprowadzenia ataków, które nie dokonują manipulacji rekomendacjami (gdyż one nie występują), czyli ataków:

- atak stały (ozn. **C**);
- oscylacja zachowania (ozn. **O**);
- niespójne zachowanie (ozn. **N**).

Warto jednak zauważyć, że w przypadku braku systemu TRM ataki oscylacji zachowania oraz niespójnego zachowania są mniej efektywne (w perspektywie wpływu na środowisko) niż atak stały i nie wywołują dodatkowych interesujących dla atakujących efektów[69]. Oscylacja zachowania jest bowiem nierzetelnym świadczeniem co którejś usługi, a niespójne zachowanie nierzetelnym świadczeniem usług dla co któregoś agenta. Wobec tego, w celu ograniczenia liczby prezentowanych wyników badań, zostaną jedynie zaprezentowane wyniki dla ataku stałego (w przypadku pozostałych ataków ich wpływ na efektywność środowiska będzie mniejszy).

[68] Możliwe jest stosowanie oddzielnie ataku oczerniania oraz wychwalania, jednak zdecydowano się na przeprowadzenie jednego badania, składającego się z tych dwóch ataków, w celu zmniejszenia liczby przeprowadzanych badań, gdyż ich mechanizm działania i wpływ na środowisko jest podobny.

[69] W przeciwieństwie do przypadku działającego systemu TRM, gdzie brak stałości w zachowaniu agenta złośliwego powoduje problemy z oceną jego rzetelności.

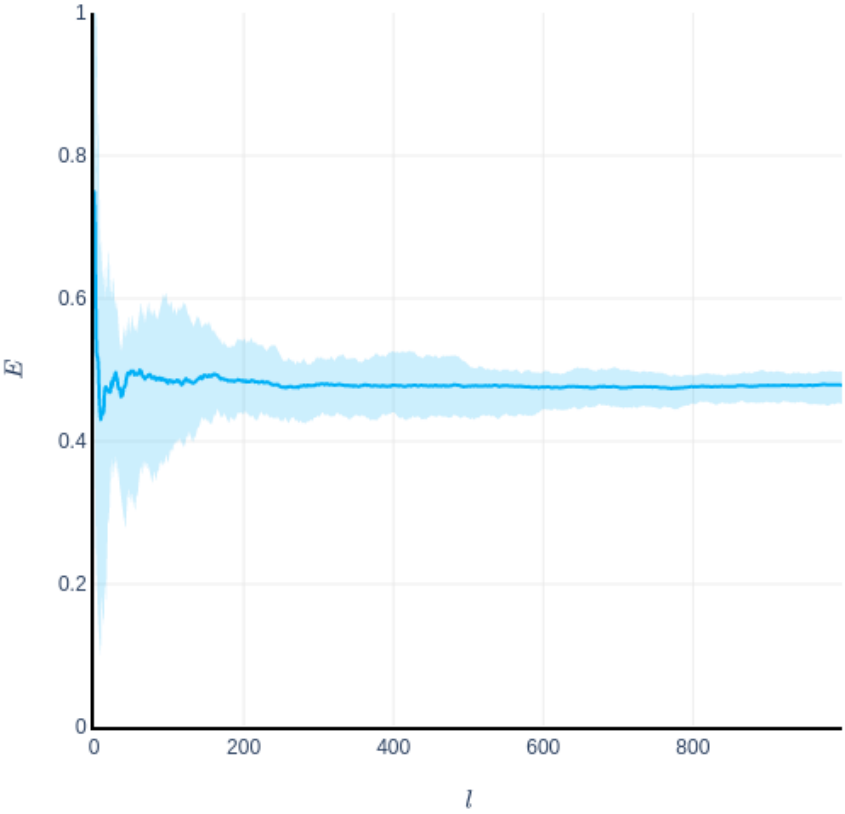
6.3.2. Badanie środowiska bez systemu TRM

W niniejszym punkcie zaprezentowano wyniki badania dla ataku stałego w przypadku braku systemu TRM.

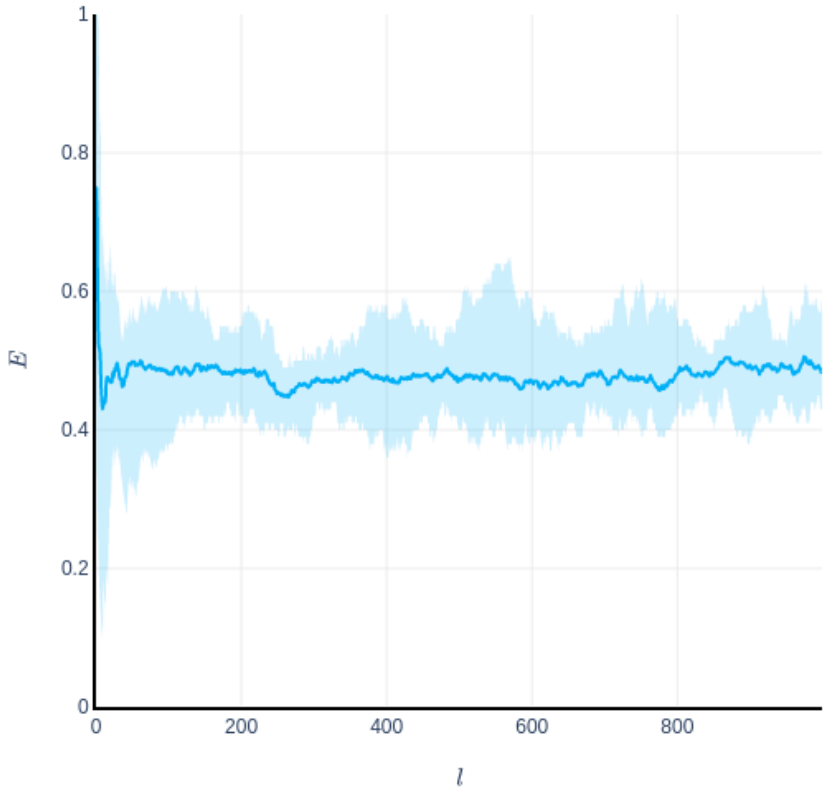
Przyjęto, że w przypadku braku systemu TRM wybór agenta jako usługodawcy dokonywany jest w sposób losowy. Wykresy prezentują wartości efektywności środowiska E (rysunek 10), efektywności chwilowej $E^{(n=100)}$ (rysunek 11) oraz liczby interakcji pomiędzy agentami rzetelnymi $l_l^{A_B,A_B}$ oraz agentami rzetelnymi i złośliwymi $l_l^{A_B,A_B}$ (rysunek 12). Tabela 5 zawiera wyniki poszczególnych miar uzyskane w tym badaniu.

Tabela 5 Wyniki badania środowiska bez systemu TRM w trakcie ataku stałego (C)

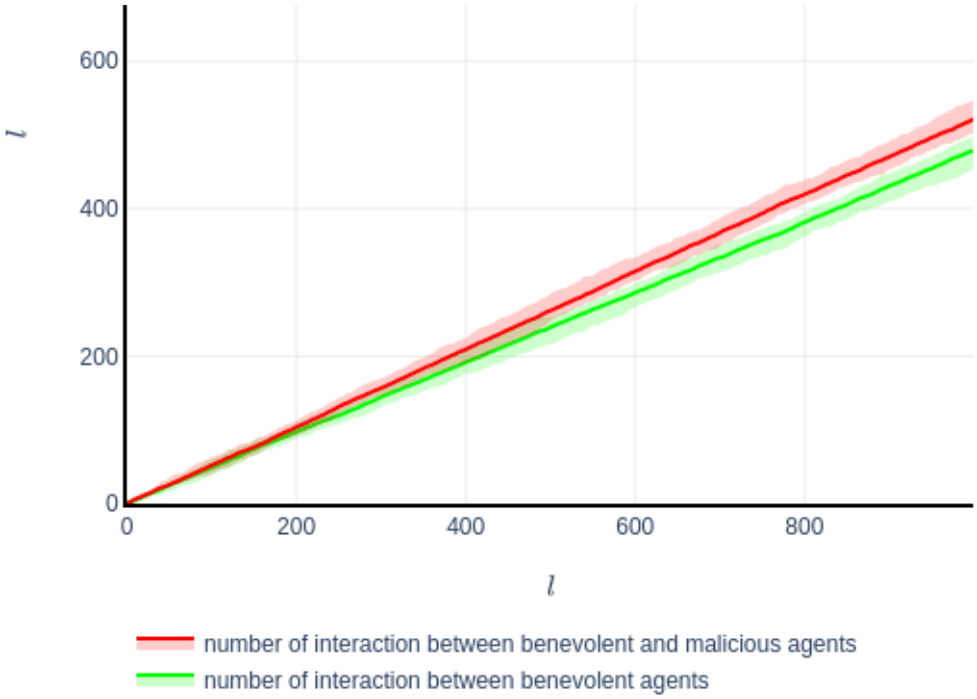
Miara	**Symbol**	**średnia**	**min**	**max**
efektywność środowiska	E	0.479	0.453	0.497
efektywność chwilowa $n = 100$	$E^{(n)}$	0.484	0.43	0.59
liczba interakcji agentów rzetelnych z agentami rzetelnymi	$l_l^{A_B,A_B}$	478.8	453.0	497.0
liczba interakcji agentów rzetelnych z agentami złośliwymi	$l_l^{A_B,A_M}$	521.2	503.0	547.0



Rysunek 10 Efektywność środowiska w przypadku braku systemu TRM w trakcie ataku stałego



Rysunek 11 Efektywność chwilowa n=100 w przypadku braku systemu TRM w trakcie ataku stałego



Rysunek 12 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku braku systemu TRM w trakcie ataku stałego

Wyniki badania pokazują, że po pewnym czasie trwania symulacji (po kilkudziesięciu interakcjach) efektywność jest bliska wartości $0{,}48$ z niewielkimi odchyleniami pomiędzy poszczególnymi przebiegami symulacji, przy czym odchylenia te są tym mniejsze im więcej interakcji zostało wykonanych. Jest to całkowicie zgodne z intuicją, ponieważ rzetelny agent może uzyskać usługę od 9 innych rzetelnych agentów (nie może wyświadczyć usługi sam sobie) oraz od 10 agentów złośliwych. Wybór każdego z agentów jest tak samo prawdopodobny, wobec czego efektywność ustala się w okolicach $\frac{9}{19} \approx 0{,}47$. Potwierdza to większa liczba interakcji rzetelnych agentów z agentami złośliwymi niż liczba interakcji pomiędzy rzetelnymi agentami.

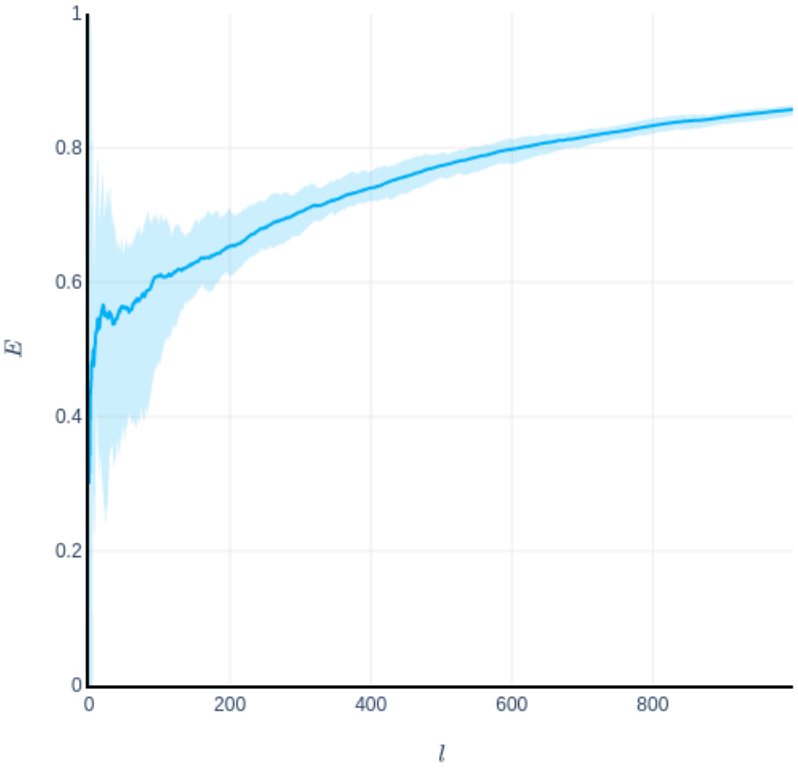
6.3.3. Badanie reakcji na ustalone ataki

Głównym celem badania jest zgrubna jakościowa ocena, czy system TRM jest podatny na dany rodzaj ataku, czy też nie.

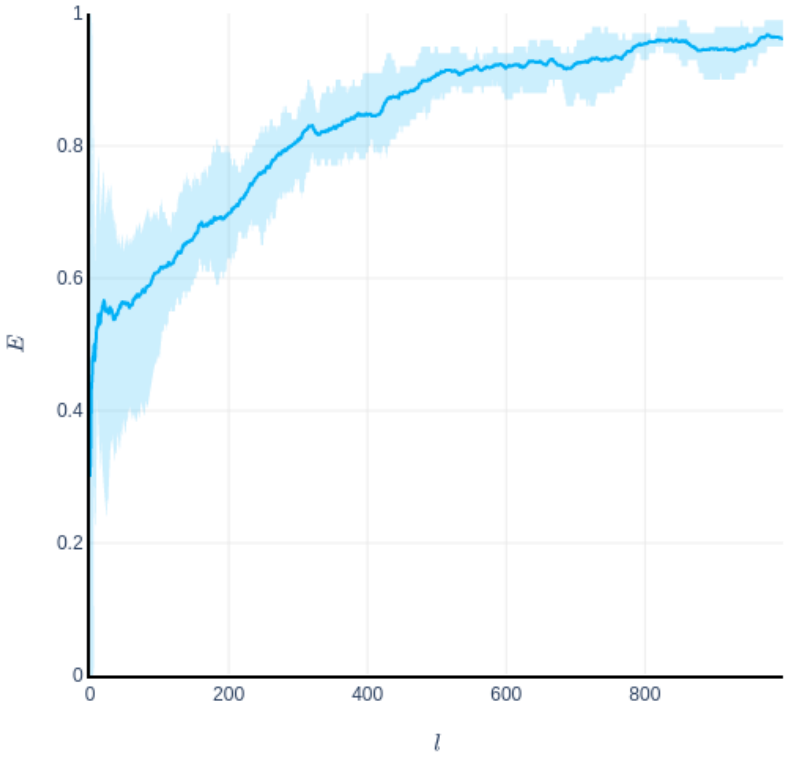
W przypadku ataków, które posiadają dodatkowe parametry, przyjęto ich wartości domyślne, określone w tabeli 2 w punkcie 6.1.4.

6.3.3.1. Atak stały

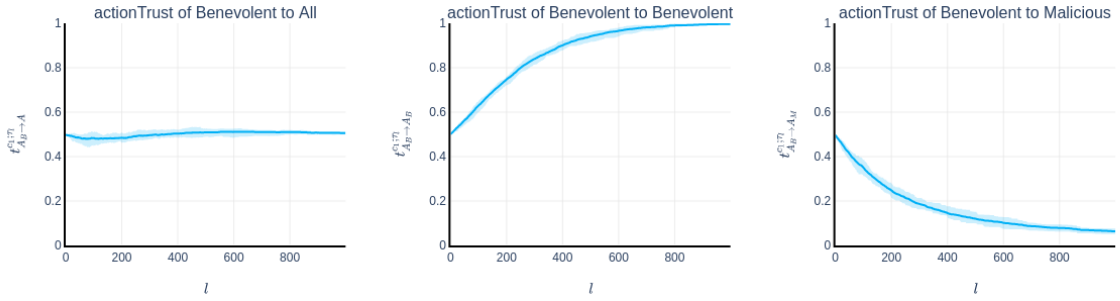
Rysunki 13–18 oraz tabela 6, prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak stały.



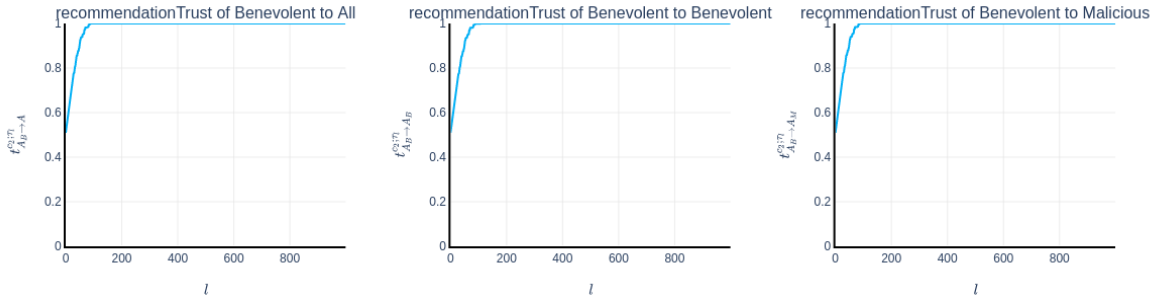
Rysunek 13 Efektywność środowiska z systemem RefTRM w trakcie ataku stałego



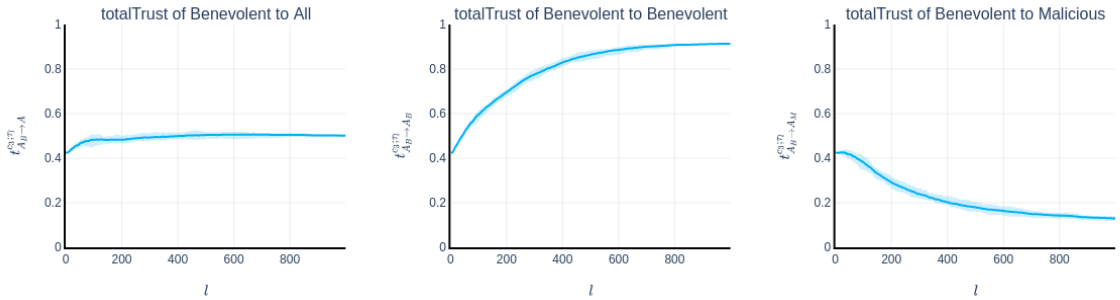
Rysunek 14 Efektywność chwilowa n=100 środowiska z systemem RefTRM w trakcie ataku stałego



Rysunek 15 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku stałego



Rysunek 16 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku stałego



Rysunek 17 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku stałego

Uzyskane wartości miar w poszczególnych momentach badania symulacyjnego nie wykazują istotnej rozbieżności z wyjątkiem parametru efektywność na początku symulacji. Jest to całkowicie zrozumiałe, gdyż na początku symulacji rzetelne agenty mają takie samo zaufanie do wszystkich pozostałych agentów, co oznacza, że dokonują wyboru usługodawcy praktycznie w losowy sposób, tak więc w zależności od przebiegu danej symulacji może się zdarzyć, że agent wybierze jako usługodawcę agenta rzetelnego lub złośliwego, co spowoduje wahania efektywności praktycznie w całym zakresie możliwych wartości tego parametru.

Rysunek 18 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku stałego

Uzyskane wyniki pokazują, że system jest częściowo podatny na ten najprostszy rodzaj ataku, ale z czasem agenty rzetelne nawiązują coraz mniej interakcji z agentami złośliwymi, o czym świadczy wzrost wartości efektywności pod koniec badania.

Warto podkreślić, że pod koniec badania zaufanie (zarówno akcyjne, jak i całkowite) do agentów złośliwych spada do minimalnej wartości osiąganej w ramach tego systemu. Wyjaśnienia wymaga, dlaczego wartość zaufania akcyjnego i całkowitego do agentów złośliwych nie osiąga 0, a jedynie dąży do 0.1. Wynika to z parametru funkcji wyboru agenta, w której prawdopodobieństwo wyboru agenta jest proporcjonalne do wartości zaufania akcyjnego, przy czym jeżeli to zaufanie jest poniżej określonego progu to na potrzeby obliczeń prawdopodobieństwa wyboru agenta, do wartości zaufania całkowitego doliczana jest dodatkowo kara (w postaci zmniejszenia prawdopodobieństwa wyboru na dostawcę usługi: β^S – parametru systemu RefTRM) w wysokości 0.1, co oznacza, że prawdopodobieństwo wyboru agenta z zaufaniem równym 0.1 lub mniej jest równe 0. Z tego względu agenty te nie są wybierane jako usługodawcy przez rzetelne agenty, co uniemożliwia dalszy spadek wartości zaufania akcyjnego do nich i w konsekwencji zaufania całkowitego.

Jeszcze bardziej zaskakujący wydaje się być fakt braku wzrostu zaufania całkowitego agentów rzetelnych do innych agentów rzetelnych – wartości te osiągają maksimum około 0.91. Wynika to z faktu, że złośliwe agenty, wskutek braku żądań świadczenia usług wysyłanych do

rzetelnych agentów (co zostało założone w badaniu), cały czas dostarczają rekomendacje o rzetelnych agentach na poziomie z początku symulacji, czyli o wartości równej 0.5. Nie jest to atak oczerniania, gdyż rzeczywiście tak te agenty oceniają agenty rzetelne. Zgodnie z parametrami systemu, jeżeli różnica pomiędzy rekomendacją (w tym przypadku 0.5) a rzeczywistym wynikiem interakcji (w przypadku agentów rzetelnych: 1.0) jest mniejsza lub równa 0.6 (a tym przypadku wynosi 0.5), to należy uznać, że rekomendacja jest prawidłowa. Świadczy też o tym wysokie zaufanie rekomendacyjne agentów rzetelnych do agentów złośliwych. Dzięki takiemu działaniu, agenty złośliwe są więc w stanie skutecznie obniżać zaufanie całkowite agentów rzetelnych do innych agentów rzetelnych, a jednocześnie nie ponosić żadnych negatywnych konsekwencji dla swoich ocen zaufania rekomendacyjnego. Sugeruje to więc, że agenty złośliwe w celu maksymalizacji skuteczności nie powinny stosować oczerniania, polegającego na maksymalnym zaniżaniu zaufania do innych agentów, a jedynie obniżać je o określoną wartość w stosunku do aktualnej jakości usługi świadczonej przez rzetelne agenty. Fakt ten, który można wszakże wysnuć także z bezpośredniej analizy systemu, zyskał potwierdzenie doświadczalne i może służyć do konstrukcji ataku dopasowanego (co zostało dokonane w punkcie 6.3.7.).

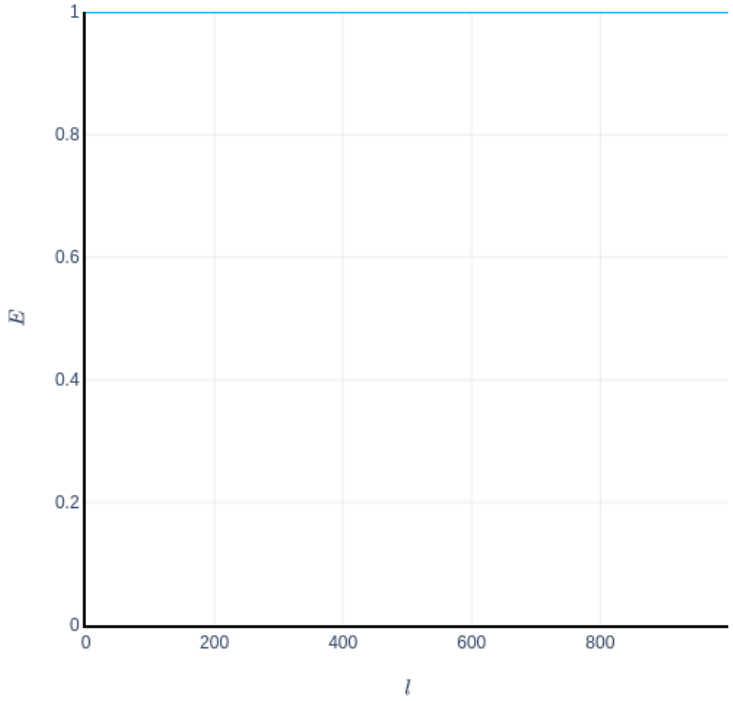
Tabela 6 Wyniki badania środowiska z systemem RefTRM w trakcie ataku stałego (C)

Miara	**Symbol**	**średnia**	**min**	**max**
efektywność środowiska	E	0.857	0.848	0.864
efektywność chwilowa n	$E^{(n)}$	0.963	0.95	0.99
globalne średnie zaufanie akcyjne agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_1;m_l}$	0.506	0.498	0.513
globalne średnie zaufanie akcyjne agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_1;m_l}$	0.997	0.991	1.0
globalne średnie zaufanie akcyjne agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_1;m_l}$	0.064	0.051	0.076
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_2;m_l}$	1.0	1.0	1.0
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_2;m_l}$	1.0	1.0	1.0
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_2;m_l}$	1.0	1.0	1.0
globalne średnie zaufanie całkowite agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_3;m_l}$	0.501	0.495	0.506
globalne średnie zaufanie całkowite agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_3;m_l}$	0.914	0.909	0.917
globalne średnie zaufanie całkowite agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_3;m_l}$	0.13	0.118	0.14
liczba interakcji agentów rzetelnych z agentami rzetelnymi	$l_I^{A_B,A_B}$	857.3	848.0	864.0
liczba interakcji agentów rzetelnych z agentami złośliwymi	$l_I^{A_B,A_M}$	142.7	136.0	152.0

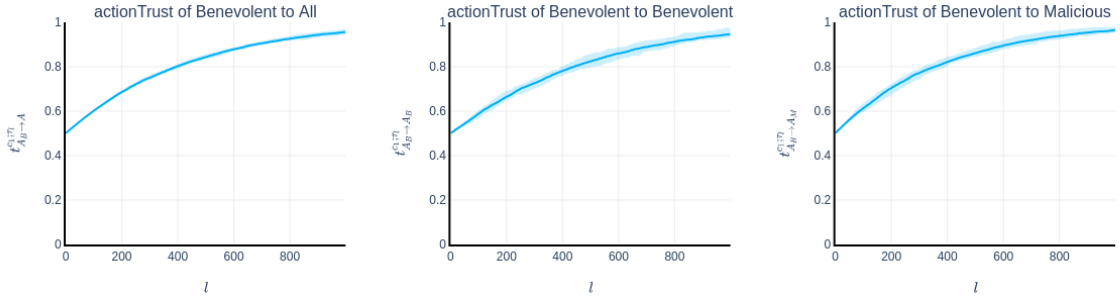
Podczas prezentacji wyników dla kolejnych ataków zostaną pominięte (w celu zaoszczędzenia miejsca) wykresy dla efektywności chwilowej o ile nie będą one znaczące dla interpretacji wyników.

6.3.3.2. Atak oczerniania i wychwalania (BF)

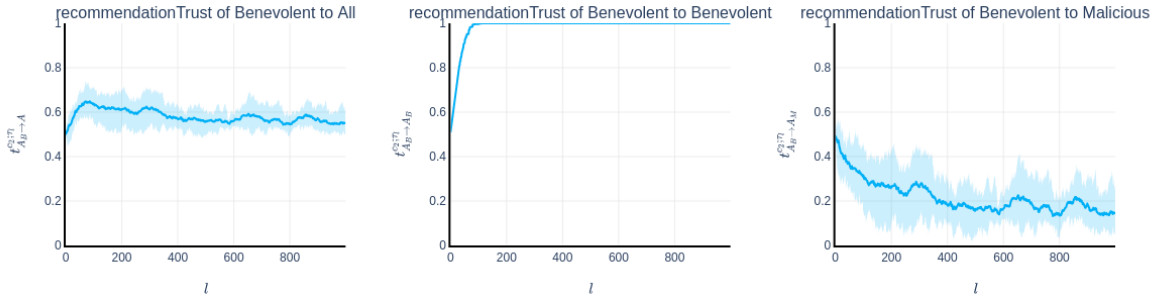
Rysunki 19–23 oraz tabela 7 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak oczerniania i wychwalania (BF).



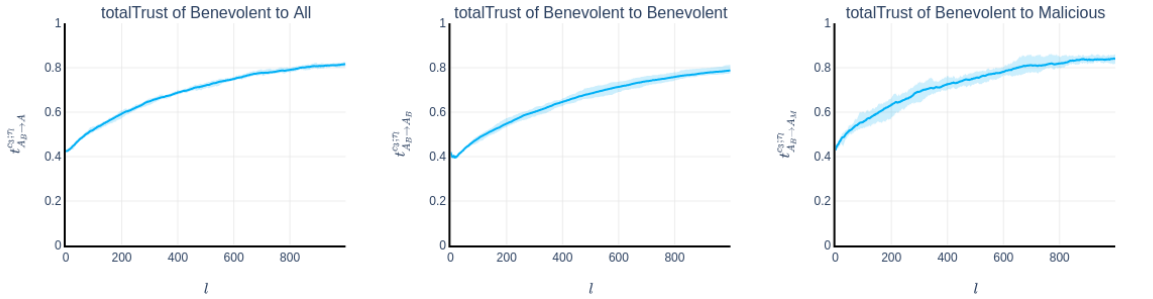
Rysunek 19 Efektywność systemu RefTRM w trakcie ataku oczerniania i wychwalania (BF)



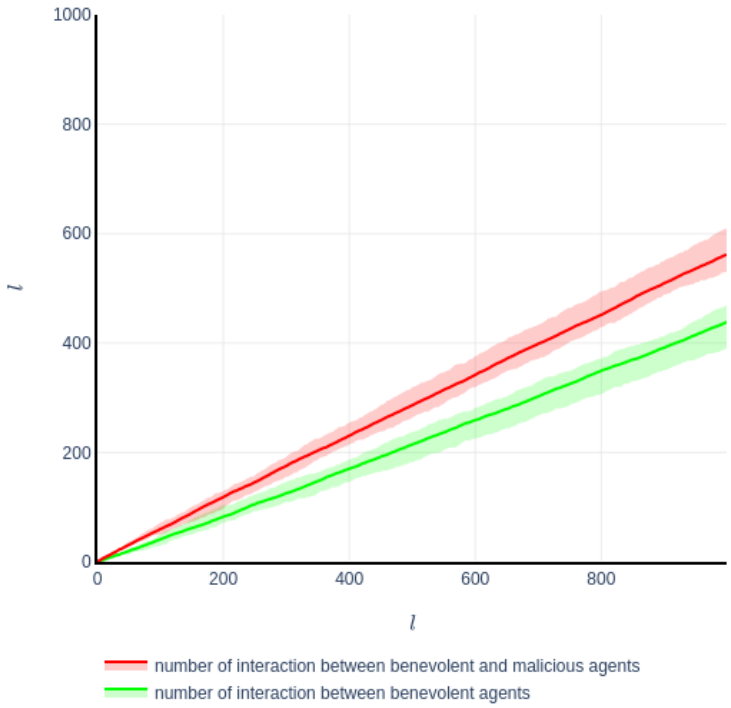
Rysunek 20 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BF



Rysunek 21 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BF



Rysunek 22 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BF



Rysunek 23 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku BF

Uzyskane wyniki pokazują, że atak ten nie ma wpływu na efektywność (jest ona równa 1 podczas całej symulacji), co nie jest w żaden sposób zaskakujące, gdyż złośliwe agenty nie zachowują się nierzetelnie podczas świadczenia usług.

Zaufanie akcyjne rzetelnych agentów zarówno do agentów rzetelnych, jak i do agentów złośliwych jest na zbliżonym poziomie podczas całej sytuacji, co jest zgodne z intuicją ze względu na to, że złośliwe agenty nie zachowują się nierzetelnie podczas interakcji, a więc agenty rzetelne nie mają powodu do obniżania zaufania akcyjnego do agentów złośliwych.

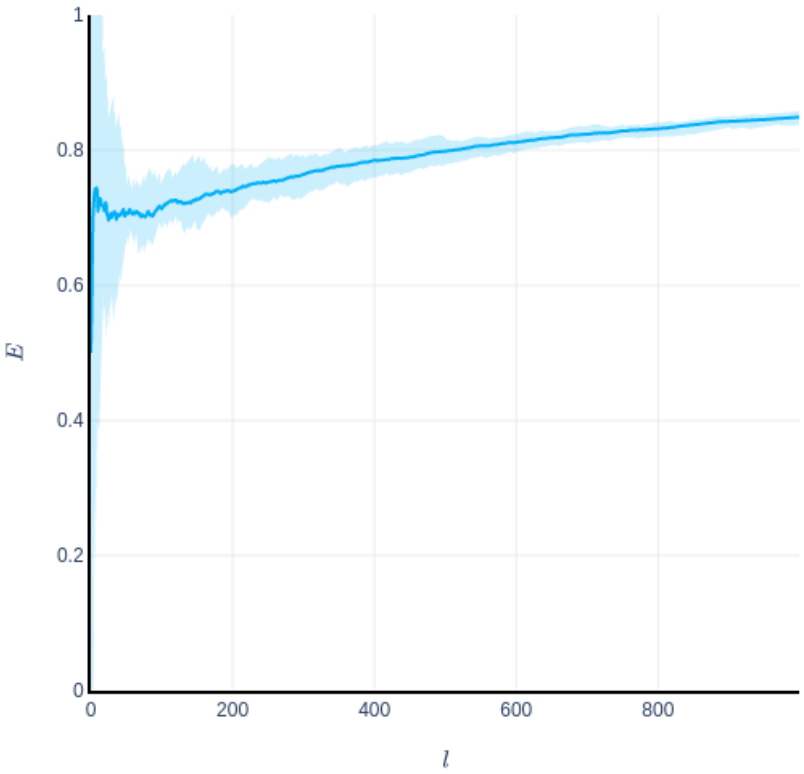
Oczerniając rzetelne agenty, atakujący nie uzyskują pożądanego efektu, tzn. nie obniżają znacząco zaufania całkowitego do agentów rzetelnych, ale powodują znaczny spadek zaufania rekomendacyjnego do nich samych.

Tabela 7 Wyniki badania środowiska z systemem RefTRM w trakcie ataku BF

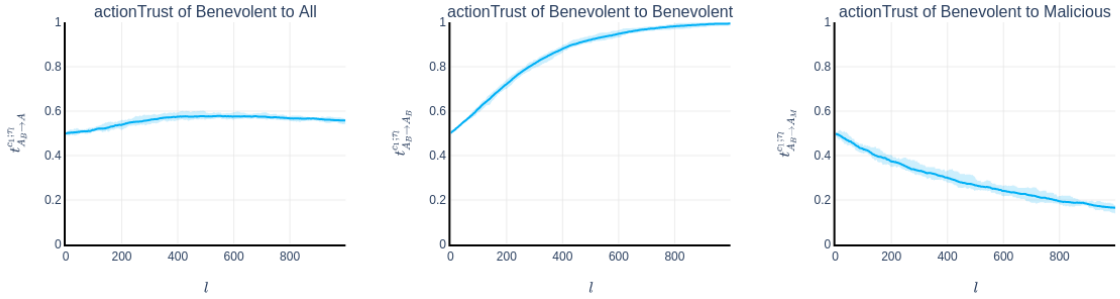
Miara	**Symbol**	**średnia**	**min**	**max**
efektywność środowiska	E	1.0	1.0	1.0
efektywność chwilowa n	$E^{(n)}$	1.0	1.0	1.0
globalne średnie zaufanie akcyjne agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_1;m_l}$	0.956	0.944	0.971
globalne średnie zaufanie akcyjne agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_1;m_l}$	0.947	0.932	0.976
globalne średnie zaufanie akcyjne agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_1;m_l}$	0.965	0.951	0.98
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_2;m_l}$	0.548	0.508	0.614
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_2;m_l}$	1.0	1.0	1.0
globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_2;m_l}$	0.142	0.066	0.267
globalne średnie zaufanie całkowite agentów rzetelnych do wszystkich agentów	$t_{A_B \to A}^{c_3;m_l}$	0.815	0.802	0.822
globalne średnie zaufanie całkowite agentów rzetelnych do agentów rzetelnych	$t_{A_B \to A_B}^{c_3;m_l}$	0.788	0.776	0.812
globalne średnie zaufanie całkowite agentów rzetelnych do agentów złośliwych	$t_{A_B \to A_M}^{c_3;m_l}$	0.84	0.821	0.859
liczba interakcji agentów rzetelnych z agentami rzetelnymi	$l_I^{A_B,A_B}$	438.4	390.0	470.0
liczba interakcji agentów rzetelnych z agentami złośliwymi	$l_I^{A_B,A_M}$	561.6	530.0	610.0

6.3.3.3. Atak oscylacji zachowania (O)

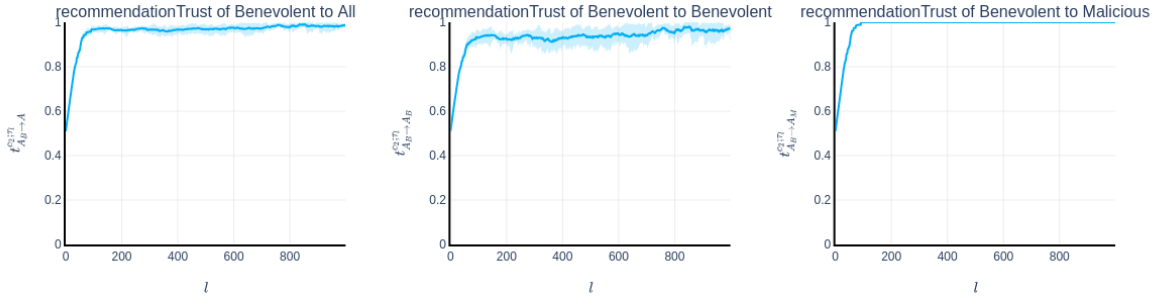
Rysunki 24–28 oraz tabela 8 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak oscylacji zachowania (O).



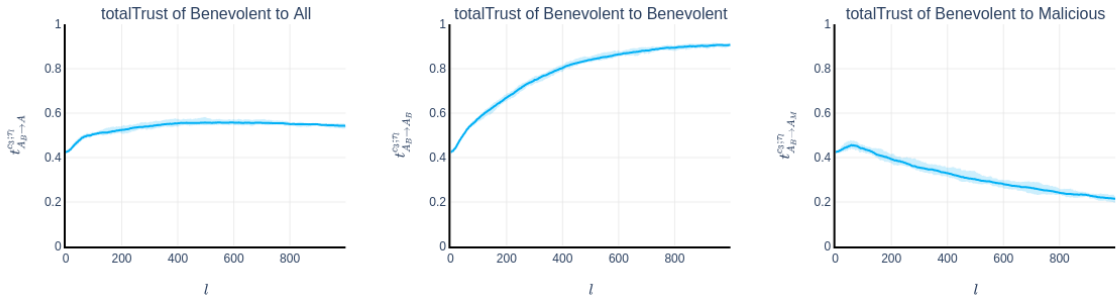
Rysunek 24 Efektywność systemu RefTRM w trakcie ataku oscylacji zachowania (O)



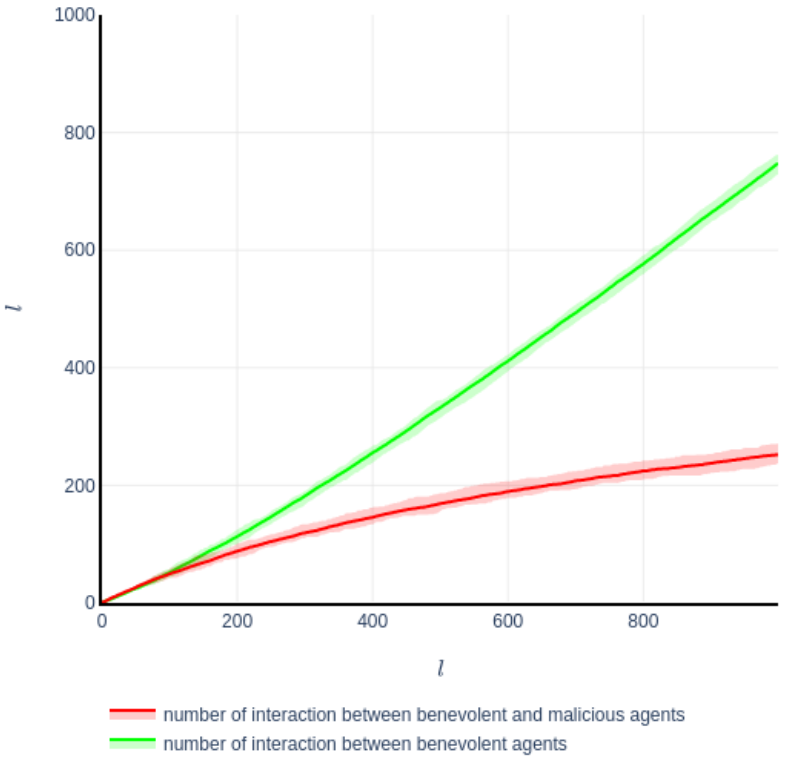
Rysunek 25 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku O



Rysunek 26 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku O



Rysunek 27 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku O



Rysunek 28 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku O

Tabela 8 Wyniki badania środowiska z systemem RefTRM w trakcie ataku O

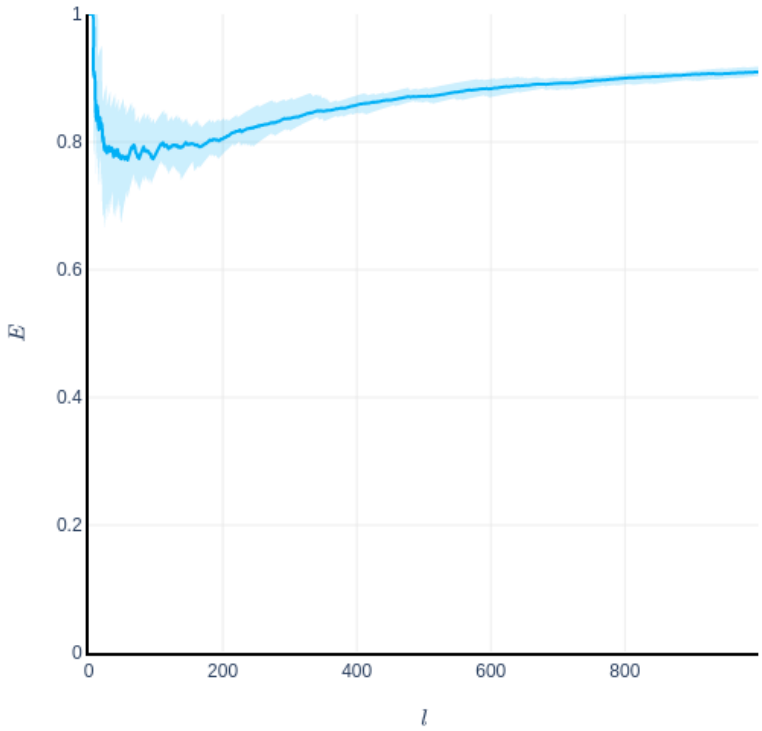
Symbol	średnia	min	max
E	0.849	0.837	0.858
$E^{(n)}$	0.911	0.85	0.95
$t_{A_B \to A}^{c_1;m_l}$	0.558	0.541	0.565
$t_{A_B \to A_B}^{c_1;m_l}$	0.994	0.986	0.999
$t_{A_B \to A_M}^{c_1;m_l}$	0.165	0.139	0.176
$t_{A_B \to A}^{c_2;m_l}$	0.985	0.96	0.996
$t_{A_B \to A_B}^{c_2;m_l}$	0.969	0.916	0.991
$t_{A_B \to A_M}^{c_2;m_l}$	1.0	1.0	1.0

$t_{A_B \to A}^{c_3;m_l}$	0.543	0.531	0.55
$t_{A_B \to A_B}^{c_3;m_l}$	0.907	0.901	0.915
$t_{A_B \to A_M}^{c_3;m_l}$	0.215	0.197	0.224
$l_l^{A_B,A_B}$	747.8	729.0	763.0
$l_l^{A_B,A_M}$	252.2	237.0	271.0

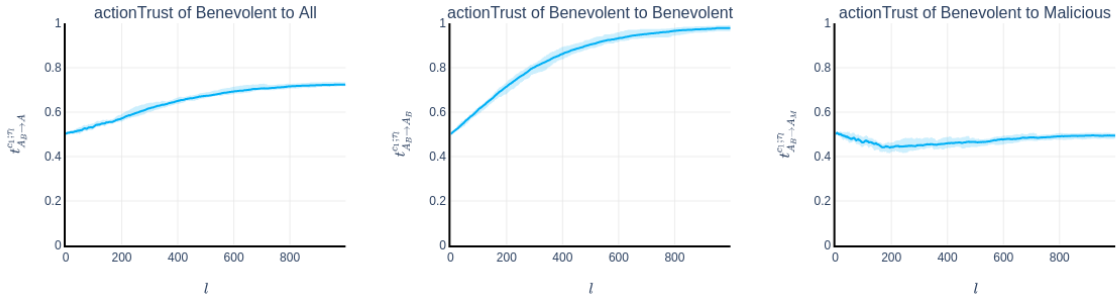
Atak oscylacji zachowania okazał się nieznacznie skuteczniejszy (powoduje osiągnięcie niższej wartości efektywności środowiska) niż atak stały. Jednak wyraźnie widać (analizując wykres efektywności środowiska, lub wartość efektywności chwilowej), że z czasem atak ten staje się coraz mniej skuteczny. Wynika to z faktu, że rzetelne agenty orientują się, że w przypadku niektórych interakcji złośliwe agenty nie dostarczają satysfakcjonującej usługi i obniżają zaufanie (akcyjne) do nich, a w konsekwencji nawiązują z agentami złośliwymi coraz mniej interakcji. Pomimo wysokiej wartości zaufania rekomendacyjnego agentów rzetelnych do agentów złośliwych (agenty złośliwe nie manipulują wszakże rekomendacjami), zaufanie całkowite jest także na niskim poziomie, co ma wpływ na wybór usługodawcy przez rzetelne agenty.

6.3.3.4. Atak niespójnego zachowania (N)

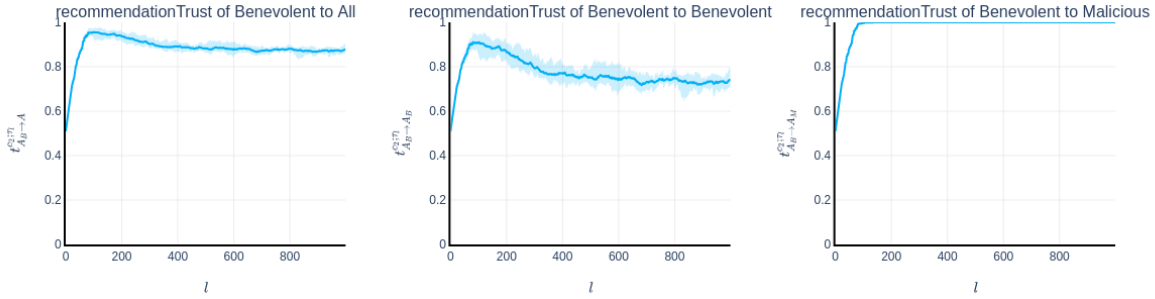
Rysunki 29–33 oraz tabela 9 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak niespójnego zachowania (N).



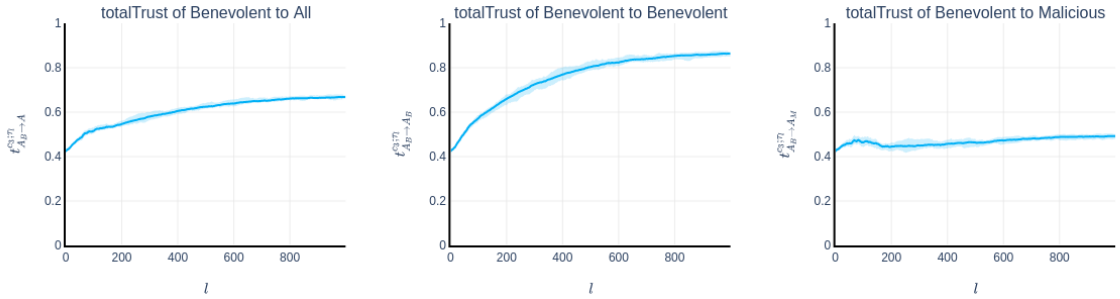
Rysunek 29 Efektywność systemu RefTRM w trakcie ataku N



Rysunek 30 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku N



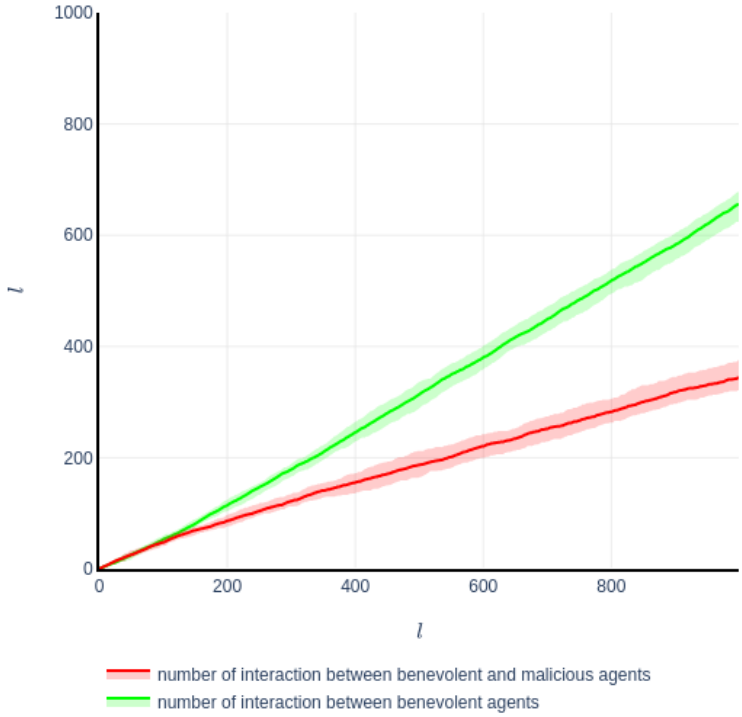
Rysunek 31 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku N



Rysunek 32 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku N

Atak ten w kontekście wpływu na efektywność środowiska jest mniej skuteczny niż atak stały i oscylacji zachowania, ale bardzo interesujące wnioski można wysnuć analizując wartość globalnego średniego zaufania rekomendacyjnego agentów rzetelnych do agentów rzetelnych. Otóż złośliwe agenty swoimi działaniami powodują, że rzetelne agenty przestają sobie ufać w kwestii rekomendacji, a zaufanie rekomendacyjne tych agentów do agentów rzetelnych jest znacznie niższe niż do agentów złośliwych. W przypadku domyślnych parametrów systemu RefTRM zaufanie rekomendacyjne ma mniejszy wpływ na zaufanie całkowite niż zaufanie akcyjne, w związku z tym, agentom złośliwym nie udaje się, skuteczniej niż w przypadku innych ataków, obniżyć efektywności środowiska. Jednak, w przypadku innych parametrów

systemu (większej wagi rekomendacji) lub w przypadku gdyby agenty rzetelne miały ograniczony zasób własnych doświadczeń, atak ten mógłby okazać się skuteczny.



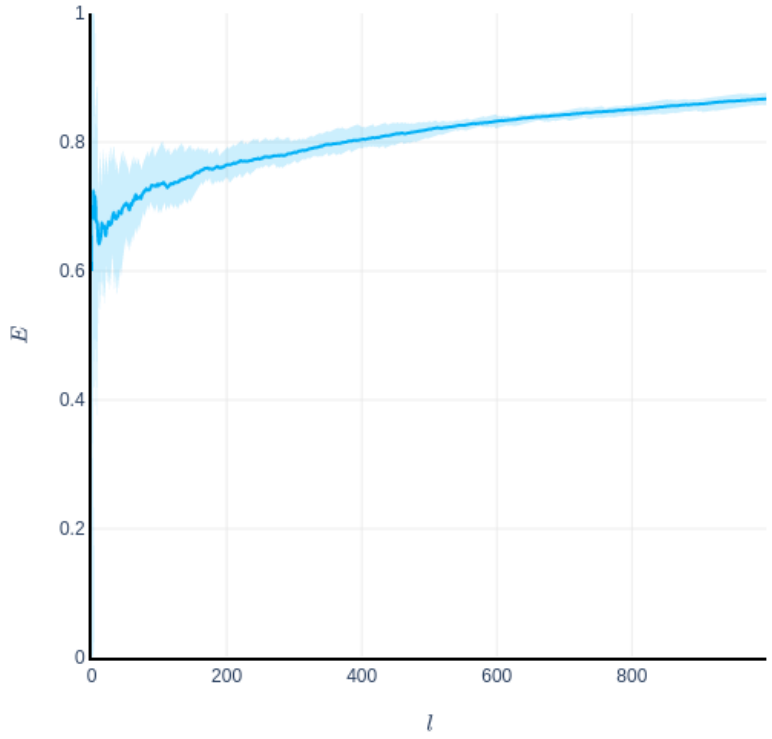
Rysunek 33 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku N

Tabela 9 Wyniki badania środowiska z systemem RefTRM w trakcie ataku N

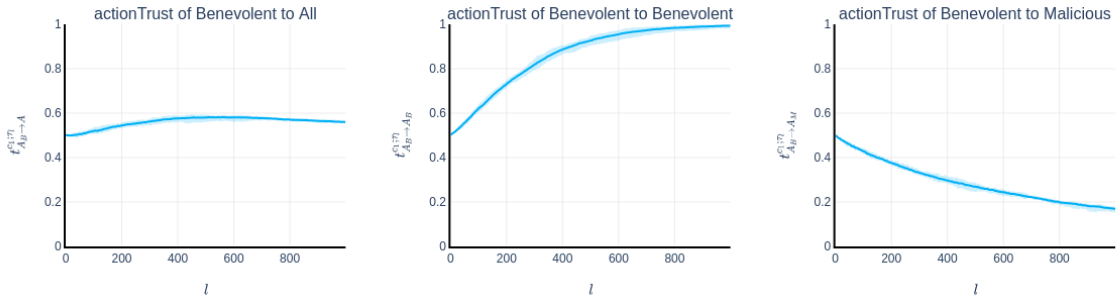
Symbol	**średnia**	**min**	**max**
E	0.91	0.903	0.918
$E^{(n)}$	0.947	0.9	0.98
$t_{A_B \to A}^{c_1;m_l}$	0.724	0.717	0.734
$t_{A_B \to A_B}^{c_1;m_l}$	0.979	0.964	0.991
$t_{A_B \to A_M}^{c_1;m_l}$	0.495	0.482	0.504
$t_{A_B \to A}^{c_2;m_l}$	0.876	0.863	0.904
$t_{A_B \to A_B}^{c_2;m_l}$	0.738	0.71	0.798
$t_{A_B \to A_M}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A}^{c_3;m_l}$	0.668	0.66	0.675
$t_{A_B \to A_B}^{c_3;m_l}$	0.864	0.85	0.872
$t_{A_B \to A_M}^{c_3;m_l}$	0.492	0.479	0.501
$l_l^{A_B,A_B}$	655.8	625.0	678.0
$l_l^{A_B,A_M}$	344.2	322.0	375.0

6.3.3.5. Atak wyrocznia (W)

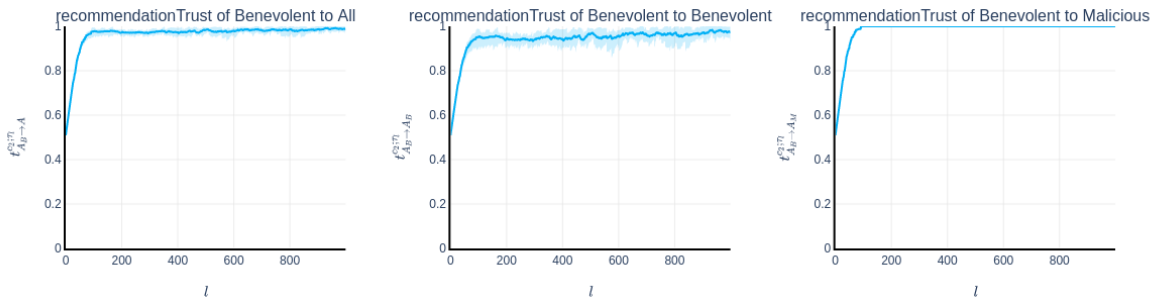
Rysunki 34–38 oraz tabela 10 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak wyrocznia (W).



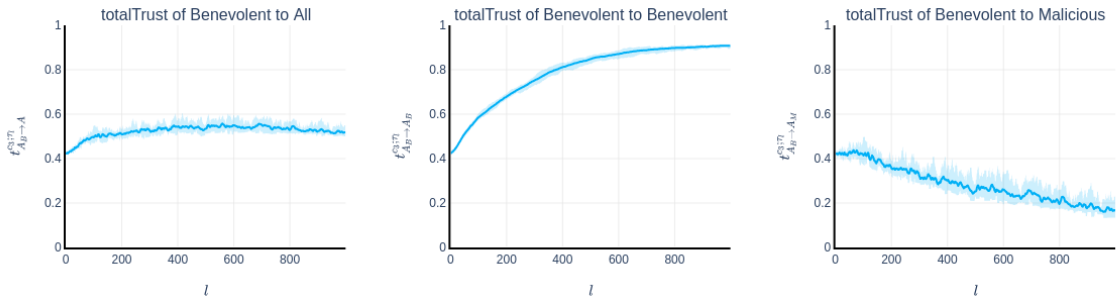
Rysunek 34 Efektywność systemu RefTRM w trakcie ataku wyrocznia (W)



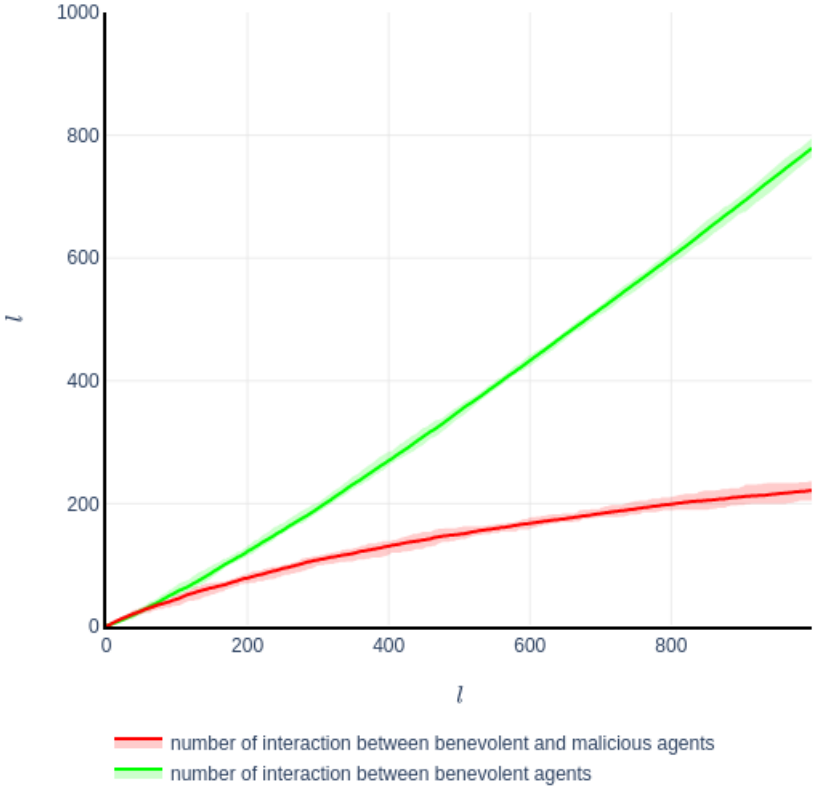
Rysunek 35 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku W



Rysunek 36 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku W



Rysunek 37 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku W



Rysunek 38 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku W

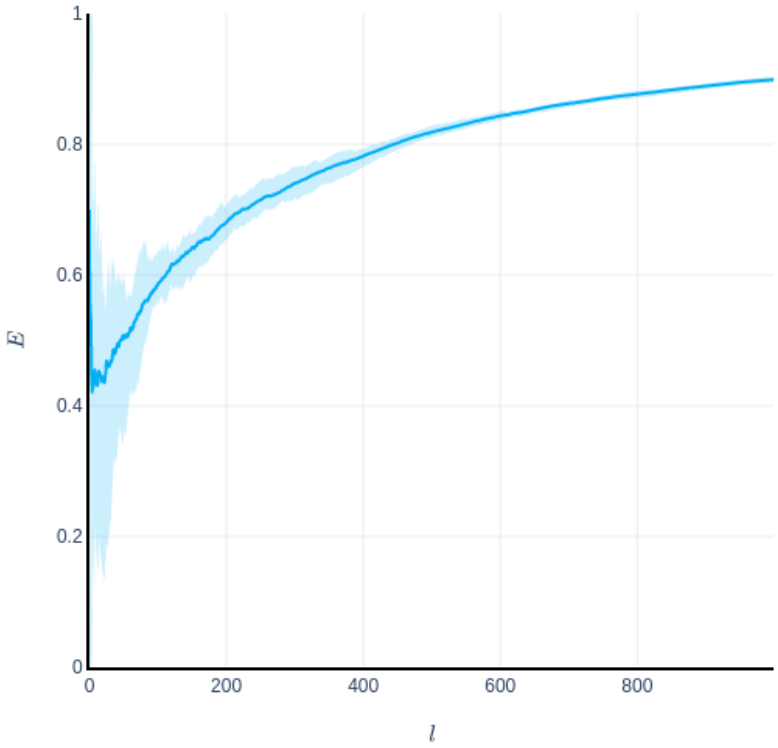
Atak wyrocznia, mimo tego, ze jest najbardziej rozbudowanym atakiem spośród analizowanych znanych ataków, gdyż zakłada kooperację złośliwych agentów i ciągłą wymianę informacji oraz ustalanie wspólnie podejmowanych działań przez złośliwe agenty, to nie wykazał się wyższą skutecznością (w odniesieniu do obniżenia efektywności środowiska) niż atak stały i atak oscylacji zachowania. Uwagę zwraca jedynie fakt, że poprzez swoje działania, atakujący zakłócają ocenę zaufania rekomendacyjnego pomiędzy rzetelnymi agentami, ale w stopniu znacznie mniejszym niż było to zaobserwowane w przypadku ataku niespójnego zachowania.

Tabela 10 Wyniki badania środowiska z systemem RefTRM w trakcie ataku W

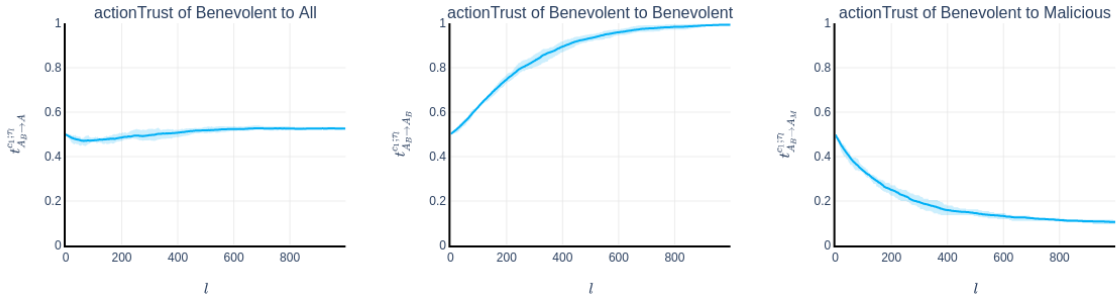
Symbol	średnia	min	max
E	0.867	0.858	0.877
$E^{(n)}$	0.94	0.92	0.97
$t_{A_B \to A}^{c_1;m_l}$	0.56	0.551	0.567
$t_{A_B \to A_B}^{c_1;m_l}$	0.994	0.984	1.0
$t_{A_B \to A_M}^{c_1;m_l}$	0.169	0.152	0.18
$t_{A_B \to A}^{c_2;m_l}$	0.988	0.976	0.999
$t_{A_B \to A_B}^{c_2;m_l}$	0.975	0.949	0.997
$t_{A_B \to A_M}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A}^{c_3;m_l}$	0.517	0.5	0.55
$t_{A_B \to A_B}^{c_3;m_l}$	0.908	0.896	0.915
$t_{A_B \to A_M}^{c_3;m_l}$	0.166	0.134	0.228
$l_l^{A_B,A_B}$	778.7	764.0	795.0
$l_l^{A_B,A_M}$	221.3	205.0	236.0

6.3.3.6. Atak oczerniania, wychwalania i stały z kooperacją (BFCc)

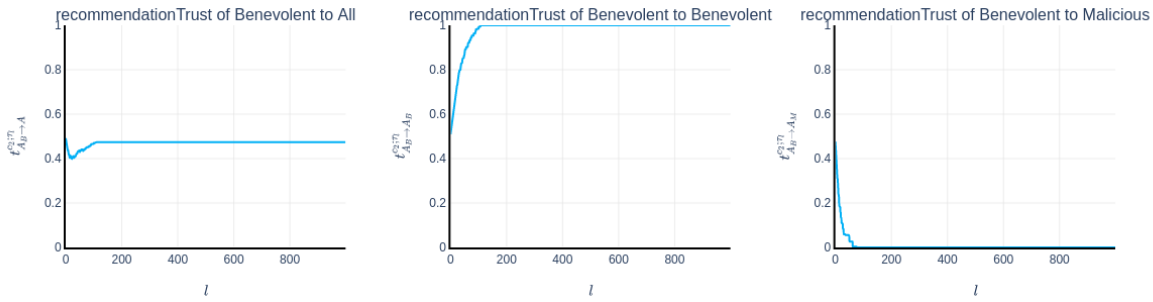
Rysunki 39–43 oraz tabela 11 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak oczerniania, wychwalania i stały z kooperacją (BFCc).



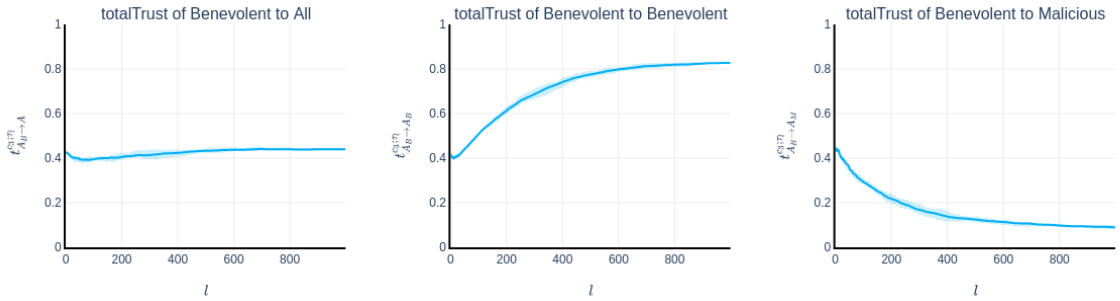
Rysunek 39 Efektywność systemu RefTRM w trakcie ataku BFCc



Rysunek 40 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFCc

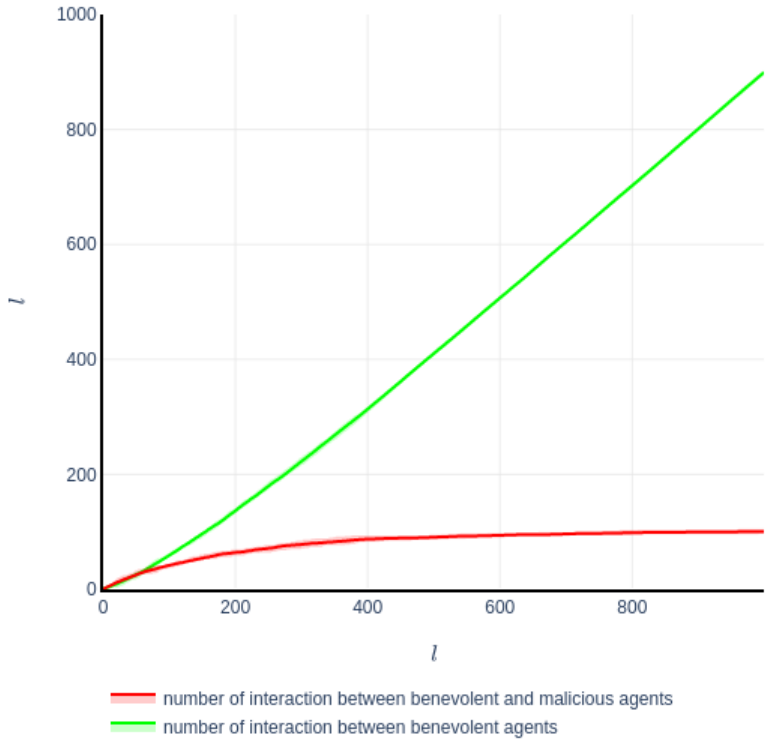


Rysunek 41 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFCc



Rysunek 42 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFCc

Podobnie jak w przypadku innych ataków uzyskane parametry w poszczególnych momentach badania symulacyjnego nie wykazują istotnej rozbieżności z wyjątkiem parametru efektywność na początku symulacji. Na podstawie uzyskanych wyników można stwierdzić, że system RefTRM jest częściowo podatny na ten rodzaj ataku, szczególnie do momentu interakcji numer 400, kiedy to efektywność środowiska jest na poziomie ok. 0.8, co oznacza, że średnio na każde 4 interakcje zakończone sukcesem, jedna kończyła się porażką (była nawiązana ze złośliwym agentem). Pod koniec symulacji, ze względu na specyfikę systemu, agenty rzetelne praktycznie już nie nawiązują interakcji z agentami złośliwymi, jednak efektywność nie jest równa 1, co wynika ze wcześniejszych interakcji, które zakończyły się brakiem świadczenia usługi.



Rysunek 43 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku BFCc

Wyniki sugerują także, że system RefTRM jest bardziej wrażliwy na wydawanie nieprawidłowych rekomendacji, niż świadczenie nierzetelnych usług (szybciej spada zaufanie rekomendacyjne do agentów złośliwych niż zaufanie akcyjne do nich). Analogicznie jest w przypadku zaufania do agentów rzetelnych, gdzie zaufanie rekomendacyjne rośnie dużo szybciej niż zaufanie akcyjne. Wynika to nie tyle z przyjętych parametrów systemu, co ze specyfiki jego funkcjonowania – agenty dużo częściej wydają rekomendacje, niż świadczą usługi, co pozwala na znacznie częstszą ocenę jakości rekomendacji niż ocenę jakości usługi.

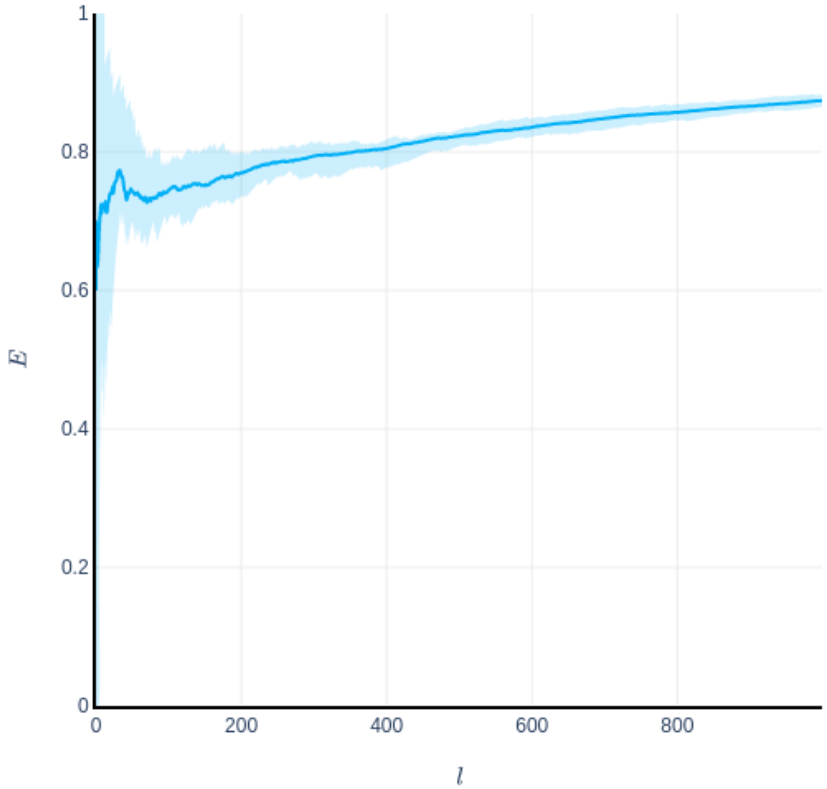
Z porównania wyników uzyskanych dla tego ataku i ataku stałego można wysnuć pozornie zaskakujący wniosek – otóż atak stały jest bardziej skuteczny przeciwko temu systemowi TRM, mimo jego znacznie mniejszego skomplikowania. O większej skuteczności ataku stałego świadczy niższa efektywność w całym przebiegu symulacji. Fakt ten jest uzasadniony przez to, że złośliwe agenty stosując atak oczerniania rzetelnych agentów i wychwalania innych złośliwych agentów, jeszcze bardziej i szybciej wpływają na zmniejszenie zaufania całkowitego do siebie. W konsekwencji, przez takie działania, zmniejszają prawdopodobieństwo wyboru złośliwych agentów jako usługodawców i tym samym zmniejszają możliwość obniżenia efektywności. Jest to uzasadnione także przez powyższą uwagę dotyczącą większej wrażliwości systemu na nieprawidłowe rekomendacje niż na nierzetelne świadczenie usług.

Tabela 11 Wyniki badania środowiska z systemem RefTRM w trakcie ataku BFCc

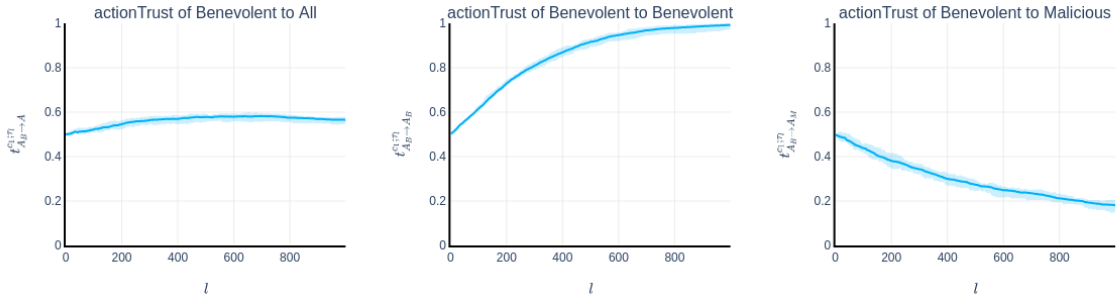
Symbol	średnia	min	max
E	0.899	0.895	0.903
$E^{(n)}$	0.99	0.98	1.0
$t_{A_B \to A}^{c_1;m_l}$	0.526	0.52	0.531
$t_{A_B \to A_B}^{c_1;m_l}$	0.994	0.989	0.998
$t_{A_B \to A_M}^{c_1;m_l}$	0.105	0.096	0.112
$t_{A_B \to A}^{c_2;m_l}$	0.474	0.474	0.474
$t_{A_B \to A_B}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A_M}^{c_2;m_l}$	0.0	0.0	0.0
$t_{A_B \to A}^{c_3;m_l}$	0.44	0.434	0.445
$t_{A_B \to A_B}^{c_3;m_l}$	0.828	0.824	0.831
$t_{A_B \to A_M}^{c_3;m_l}$	0.09	0.082	0.098
$l_l^{A_B,A_B}$	899.2	895.0	903.0
$l_l^{A_B,A_M}$	100.8	97.0	105.0

6.3.3.7. Atak oczerniania, wychwalania i oscylacji zachowania z kooperacją (BFOc)

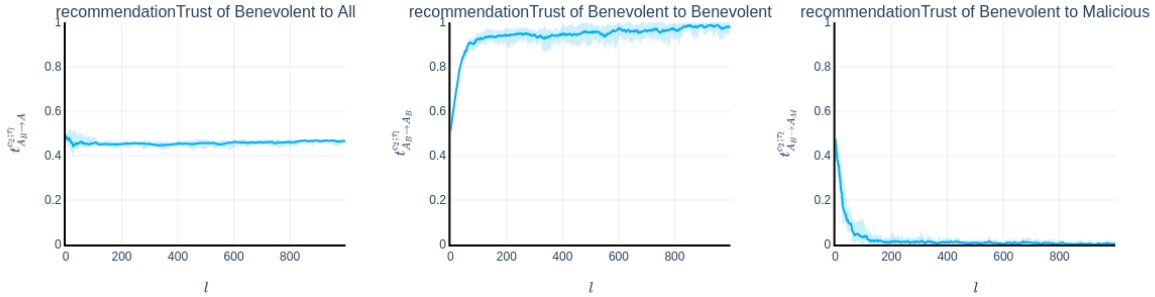
Rysunki 44–48 oraz tabela 12 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak oczerniania, wychwalania i oscylacji zachowania z kooperacją (BFOc).



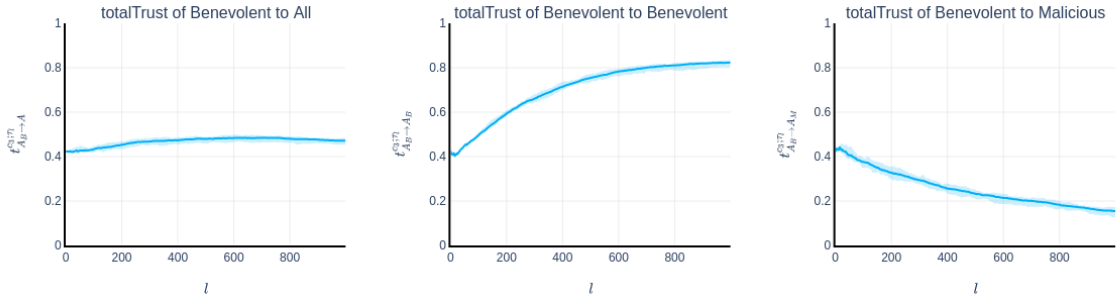
Rysunek 44 Efektywność systemu RefTRM w trakcie ataku BFOc



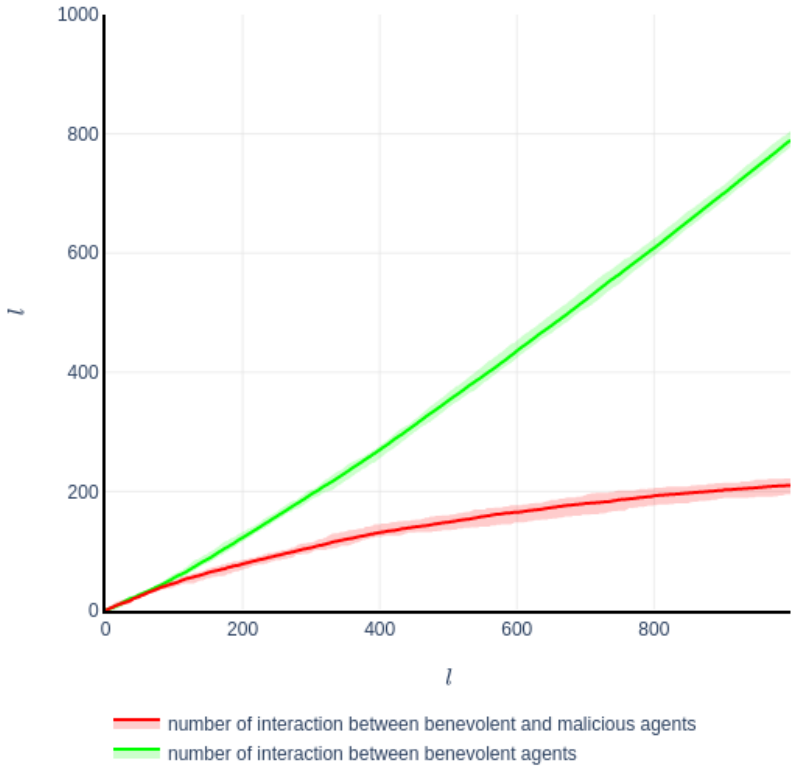
Rysunek 45 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFOc



Rysunek 46 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFOc



Rysunek 47 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku BFOc



Rysunek 48 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku BFOc

Tabela 12 Wyniki badania środowiska z systemem RefTRM w trakcie ataku BFOc

Symbol	**średnia**	**min**	**max**
E	0.874	0.865	0.883
$E^{(n)}$	0.947	0.92	0.98
$t_{A_B \to A}^{c_1;m_l}$	0.565	0.547	0.579
$t_{A_B \to A_B}^{c_1;m_l}$	0.992	0.972	1.0
$t_{A_B \to A_M}^{c_1;m_l}$	0.181	0.15	0.204
$t_{A_B \to A}^{c_2;m_l}$	0.466	0.456	0.473
$t_{A_B \to A_B}^{c_2;m_l}$	0.98	0.951	0.998
$t_{A_B \to A_M}^{c_2;m_l}$	0.005	0.0	0.01
$t_{A_B \to A}^{c_3;m_l}$	0.472	0.454	0.483
$t_{A_B \to A_B}^{c_3;m_l}$	0.824	0.805	0.833
$t_{A_B \to A_M}^{c_3;m_l}$	0.154	0.126	0.173
$l_I^{A_B,A_B}$	789.6	778.0	805.0
$l_I^{A_B,A_M}$	210.4	195.0	222.0

Atak BFOc okazał się skuteczniejszy niż atak BFCc, ale mniej skuteczny niż atak oscylacji zachowania. W przypadku tego systemu TRM i domyślnych parametrów systemu, manipulowanie rekomendacjami przez złośliwe agenty nie przynosi im dodatkowych korzyści, co więcej, powoduje bardzo istotny spadek zaufania rekomendacyjnego do takich agentów.

6.3.3.8. Podsumowanie

W odniesieniu do sposobu przeprowadzenia badań warto poczynić dwie istotne uwagi. Z jednej strony prawdopodobne jest osiągnięcie większej skuteczności systemu dla zaprezentowanych ataków poprzez zmianę parametrów systemu (np. zwiększenie wagi zaufania akcyjnego). Z drugiej strony możliwe są warianty bardziej skutecznych ataków (np. poprzez ograniczenie zmiany wartości rekomendacji w trakcie ataków wychwalania i oczerniania).

Istotnym wnioskiem z przeprowadzonych badań jest to, że wyniki pozwalają dostrzec właściwości systemu RefTRM, które jasno nie wynikają z bezpośredniej analizy systemu.

W tabelach 13, 15, 16 zaprezentowano jedynie średnie wartości poszczególnych miar obliczone z wszystkich przebiegów badania, a nie zaprezentowano wartości minimalnych i maksymalnych, jak w powyższych tabelach.

Tabela 13 Zestawienie wyników badania środowiska z systemem RefTRM oraz bez systemu podczas różnych ataków

Symbol	**Bez TRM atak C**	**RefTRM atak C**	**RefTRM atak O**	**RefTRM atak BFOc**	**RefTRM atak BFCc**	**RefTRM atak W**	**RefTRM atak N**	**RefTRM atak BF**
E	0.479	0.857	0.849	0.874	0.899	0.867	0.91	1.0
$E^{(n)}$	0.484	0.963	0.911	0.947	0.99	0.94	0.947	1.0
$t_{A_B \to A}^{c_1;m_l}$		0.506	0.558	0.565	0.526	0.56	0.724	0.956
$t_{A_B \to A_B}^{c_1;m_l}$		0.997	0.994	0.992	0.994	0.994	0.979	0.947
$t_{A_B \to A_M}^{c_1;m_l}$		0.064	0.165	0.181	0.105	0.169	0.495	0.965
$t_{A_B \to A}^{c_2;m_l}$		1.0	0.985	0.466	0.474	0.988	0.876	0.548
$t_{A_B \to A_B}^{c_2;m_l}$		1.0	0.969	0.98	1.0	0.975	0.738	1.0
$t_{A_B \to A_M}^{c_2;m_l}$		1.0	1.0	0.005	0.0	1.0	1.0	0.142
$t_{A_B \to A}^{c_3;m_l}$		0.501	0.543	0.472	0.44	0.517	0.668	0.815
$t_{A_B \to A_B}^{c_3;m_l}$		0.914	0.907	0.824	0.828	0.908	0.864	0.788
$t_{A_B \to A_M}^{c_3;m_l}$		0.13	0.215	0.154	0.09	0.166	0.492	0.84
$l_I^{A_B,A_B}$	478.8	857.3	747.8	789.6	899.2	778.7	655.8	438.4
$l_I^{A_B,A_M}$	521.2	142.7	252.2	210.4	100.8	221.3	344.2	561.6

Atak oscylacji zachowania okazał się najskuteczniejszy (spośród dobrze znanych ataków) dla środowiska poddanego badaniom, w przypadku wykorzystania przez niego systemu RefTRM. Warto jednak zaznaczyć, że im dłużej będzie funkcjonować środowisko tym skuteczność tego ataku będzie spadać. Świadczy o tym zarówno wyższa wartość efektywności chwilowej pod koniec badania od efektywności środowiska oznaczonej podczas całego badania, jak i niskie wartości globalnego średniego zaufania akcyjnego i całkowitego agentów

rzetelnych do agentów złośliwych, co spowoduje, że agenty te będą rzadko wybierane przez rzetelne agenty jako usługodawcy.

6.3.4. Badanie wpływu wartości parametrów systemu TRM

Niniejszy punkt zawiera porównanie uzyskiwanych wyników badań dla różnych parametrów systemu TRM.

Tabela 14 Wartości parametrów systemu RefTRM: domyślne oraz przyjęte w badaniach

Symbol	Nazwa lub opis parametru	Wartość domyślna	Wartości w badaniach
$t_{init}^{c_1}$	początkowe zaufanie akcyjne	0.5	0.5
$t_{init}^{c_2}$	początkowe zaufanie rekomendacyjne	0.5	0.5
α^A	zwiększenie zaufania akcyjnego za satysfakcjonującą jakość usługi	0.2	0.05; 0.2; 0.4
β^A	zmniejszenie zaufania akcyjnego za niesatysfakcjonującą jakość usługi	0.4	0.1; 0.4; 0.8
α^R	zwiększenie zaufania rekomendacyjnego za prawidłową rekomendację	0.1	0.02; 0.1; 0.2
β^R	zmniejszenie zaufania rekomendacyjnego za nieprawidłową rekomendację	0.25	0.05; 0.25; 0.5
h	próg minimalnego zaufania	0.5	0.2; 0.5; 0.8
h^A	próg satysfakcjonującej usługi	0.5	0.2; 0.5; 0.8
h^R	próg prawidłowości rekomendacji	0.6	0.2; 0.6; 0.8
α	waga zaufania akcyjnego	0.7	0; 0.5; 0.7; 1
β^S	zmniejszenie prawdopodobieństwa wyboru na dostawcę usługi	0.1	0.1; 0.5; 1

Zdecydowano się nie uwzględnić zmian początkowych wartości zaufania akcyjnego i rekomendacyjnego (odpowiednio: $t_{init}^{c_1}$ i $t_{init}^{c_2}$), ze względu na to, że ich zmiany miałyby wpływ na uzyskiwane wyniki jedynie w początkowym okresie badania.

Badania wykonano zmieniając pojedynczy parametr w stosunku do wartości domyślnych. Taki sposób był spowodowany faktem, że sprawdzenie każdej kombinacji 9 lub 11 parametrów, nawet dla kilku możliwych wartości parametrów, wygenerowałoby zbyt dużą liczbę badań. Wartości parametrów przyjęte w poszczególnych badaniach zawarto w tabeli 14, a wyniki badań dla różnych ataków w tabeli 15.

Tabela 15 Wyniki badań efektywności systemu RefTRM w zależności od jego parametrów podczas różnych ataków

	Parametry systemu RefTRM										Miary efek.	
Badanie	α^A	β^A	α^R	β^R	h	h^A	h^R	α	β^S	atak	E	$E^{(n)}$
1	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
2	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
3	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
4	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0

	Parametry systemu RefTRM										Miary efek.	
Badanie	α^A	β^A	α^R	β^R	h	h^A	h^R	α	β^S	atak	E	$E^{(n)}$
5	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	BF	1.0	1.0
6	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	BF	1.0	1.0
7	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	BF	1.0	1.0
8	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	BF	1.0	1.0
9	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	BF	1.0	1.0
10	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
11	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	BF	1.0	1.0
12	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	BF	1.0	1.0
13	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	BF	1.0	1.0
14	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	BF	1.0	1.0
15	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	BF	1.0	1.0
16	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	BF	1.0	1.0
17	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
18	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
19	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
20	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BF	1.0	1.0
21	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.9	1.0
22	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.771	0.83
23	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.9	0.98
24	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	BFC	0.897	1.0
25	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	BFC	0.899	0.99
26	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	BFC	0.898	1.0
27	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	BFC	0.901	0.99
28	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	BFC	0.865	0.94
29	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	BFC	0.897	0.99
30	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.899	0.98
31	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	BFC	0.902	1.0
32	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	BFC	0.899	0.99
33	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	BFC	0.903	1.0
34	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	BFC	0.898	0.98
35	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	BFC	0.899	0.99
36	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	BFC	0.898	0.99
37	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	BFC	0.901	0.99
38	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.899	0.99
39	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.9	0.98
40	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFC	0.904	1.0
41	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.886	0.97
42	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.761	0.83
43	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.878	0.91
44	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	BFO	0.874	0.97
45	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	BFO	0.861	0.92
46	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	BFO	0.873	0.96
47	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	BFO	0.875	0.98
48	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	BFO	0.851	0.93
49	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	BFO	0.87	0.89
50	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.86	0.94

	Parametry systemu RefTRM										Miary efek.	
Badanie	α^A	β^A	α^R	β^R	h	h^A	h^R	α	β^S	atak	E	$E^{(n)}$
51	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	BFO	0.864	0.97
52	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	BFO	0.86	0.95
53	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	BFO	0.874	0.98
54	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	BFO	0.869	0.92
55	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	BFO	0.871	0.95
56	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	BFO	0.874	0.93
57	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	BFO	0.869	0.93
58	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.86	0.92
59	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.907	0.99
60	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	BFO	0.845	0.91
61	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	C	0.849	0.95
62	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	C	0.717	0.85
63	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	C	0.86	0.97
64	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	C	0.855	0.97
65	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	C	0.855	0.93
66	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	C	0.856	0.97
67	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	C	0.901	0.99
68	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	C	0.741	0.89
69	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	C	0.827	0.96
70	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	C	0.859	0.95
71	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	C	0.864	0.96
72	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	C	0.862	0.96
73	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	C	0.902	0.99
74	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	C	0.852	0.95
75	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	C	0.857	0.94
76	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	C	0.859	0.97
77	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	C	0.858	0.95
78	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	C	0.85	0.92
79	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	C	0.897	0.98
80	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	C	0.861	0.99
81	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	N	0.903	0.93
82	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	N	0.852	0.87
83	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	N	0.919	0.98
84	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	N	0.908	0.98
85	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	N	0.892	0.95
86	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	N	0.917	0.95
87	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	N	0.947	0.98
88	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	N	0.808	0.9
89	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	N	0.889	0.94
90	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	N	0.906	0.96
91	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	N	0.922	0.98
92	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	N	0.911	0.95
93	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	N	0.95	1.0
94	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	N	0.911	0.94
95	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	N	0.909	0.96
96	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	N	0.909	0.94

	Parametry systemu RefTRM										Miary efek.	
Badanie	α^A	β^A	α^R	β^R	h	h^A	h^R	α	β^S	atak	E	$E^{(n)}$
97	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	N	0.921	0.97
98	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	N	0.912	0.96
99	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	N	0.925	0.98
100	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	N	0.908	0.93
101	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	O	0.86	0.91
102	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	O	0.746	0.77
103	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	O	0.85	0.91
104	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	O	0.848	0.9
105	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	O	0.836	0.89
106	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	O	0.845	0.89
107	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	O	0.873	0.9
108	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	O	0.789	0.8
109	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	O	0.836	0.91
110	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	O	0.858	0.95
111	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	O	0.831	0.93
112	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	O	0.855	0.87
113	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	O	0.874	0.93
114	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	O	0.85	0.92
115	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	O	0.851	0.96
116	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	O	0.847	0.93
117	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	O	0.856	0.91
118	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	O	0.862	0.93
119	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	O	0.898	0.96
120	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	O	0.836	0.83
121	0.05	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	W	0.913	0.93
122	0.2	0.1	0.1	0.25	0.5	0.5	0.6	0.7	0.1	W	0.896	0.89
123	0.2	0.4	0.02	0.25	0.5	0.5	0.6	0.7	0.1	W	0.905	0.91
124	0.2	0.4	0.1	0.05	0.5	0.5	0.6	0.7	0.1	W	0.907	0.91
125	0.2	0.4	0.1	0.25	0.2	0.5	0.6	0.7	0.1	W	0.903	0.9
126	0.2	0.4	0.1	0.25	0.5	0.2	0.6	0.7	0.1	W	0.909	0.92
127	0.2	0.4	0.1	0.25	0.5	0.5	0.2	0.7	0.1	W	0.91	0.92
128	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.0	0.1	W	0.903	0.9
129	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.5	0.1	W	0.907	0.91
130	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	W	0.907	0.91
131	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.5	W	0.9	0.9
132	0.2	0.4	0.1	0.25	0.5	0.5	0.6	0.7	1.0	W	0.905	0.91
133	0.2	0.4	0.1	0.25	0.5	0.5	0.6	1.0	0.1	W	0.913	0.91
134	0.2	0.4	0.1	0.25	0.5	0.5	0.8	0.7	0.1	W	0.904	0.89
135	0.2	0.4	0.1	0.25	0.5	0.8	0.6	0.7	0.1	W	0.902	0.91
136	0.2	0.4	0.1	0.25	0.8	0.5	0.6	0.7	0.1	W	0.912	0.92
137	0.2	0.4	0.1	0.5	0.5	0.5	0.6	0.7	0.1	W	0.901	0.91
138	0.2	0.4	0.2	0.25	0.5	0.5	0.6	0.7	0.1	W	0.907	0.89
139	0.2	0.8	0.1	0.25	0.5	0.5	0.6	0.7	0.1	W	0.914	0.92
140	0.4	0.4	0.1	0.25	0.5	0.5	0.6	0.7	0.1	W	0.903	0.89

W tabeli 15 kolorem żółtym zaznaczono badania, w których przyjęto domyślne parametry systemu RefTRM w przypadku stosowania poszczególnych ataków. Kolorem czerwonym zaznaczono najniższe wartości efektywności i efektywności chwilowej w badaniach poszczególnych ataków. Kolorem zielonym zaznaczono najwyższe wartości efektywności i efektywności chwilowej w badaniach poszczególnych ataków. Atak BF nie wpływa na efektywność, dlatego wartości efektywności w jego przypadku nie zostały wyróżnione.

Analiza wyników ujawnia, że im większa wartość kary za nierzetelną usługę tym większa odporność systemu RefTRM na ataki. Wniosek ten nie jest zaskakujący w przypadku tego środowiska. Można domniemywać, że najbardziej efektywny system TRM w przypadku tego środowiska zakładałby następujące działanie rzetelnych agentów: jeżeli dany agent wyświadczył złą usługę to dodaj go do czarnej listy i nie żądaj od niego więcej usług. Wtedy każdy złośliwy agent mógłby wyświadczyć, co najwyżej jedną złą usługę każdemu rzetelnemu agentowi. W przypadku systemu RefTRM takie działanie odpowiadałoby wartościom parametrów: $\beta^A = 1$, $\beta^S = 1$, $\alpha = 1$. Możliwe byłoby także wykorzystanie rekomendacji, co dodatkowo mogłoby zwiększyć odporność systemu RefTRM na ataki (zwiększyć efektywność w przypadku ataków). Warto jednakże zauważyć, że takie działanie przestałoby być efektywne w przypadku gdy w środowisku występowałyby zakłócenia.

6.3.5. Badanie wpływu doboru parametrów ataków

Punkt zawiera badanie wpływu liczby złośliwych agentów i parametrów ataku na efektywność środowiska. Parametry ataków przyjęto zgodnie z tabelą 2 w punkcie 6.1.4, przy czym przeprowadzono badania w przypadku 2, 5, 10 i 15 złośliwych agentów funkcjonujących w środowisku. Wyniki badań zostały zaprezentowane w tabeli 16.

Tabela 16 Badania efektywności systemu RefTRM podczas ataków o różnych parametrach wykonywanych przez różną liczbę złośliwych agentów

	Parametry ataku			**Miary efek.**	
Badanie	n_M	atak	parametry ataku	E	$E^{(n)}$
0	2	BF	nd.	1.0	1.0
1	2	BFC	nd.	0.972	1.0
2	2	BFO	{'ratio': 0.1, 'interval': 10}	0.968	1.0
3	2	BFO	{'ratio': 0.2, 'interval': 5}	0.97	1.0
4	2	BFO	{'ratio': 0.4, 'interval': 5}	0.975	0.99
5	2	BFO	{'ratio': 0.6, 'interval': 5}	0.974	0.96
6	2	BFO	{'ratio': 0.8, 'interval': 5}	0.976	0.98
7	2	C	nd.	0.963	1.0

	Parametry ataku			Miary efek.	
Badanie	n_M	atak	parametry ataku	E	$E^{(n)}$
8	2	N	{'list_of_agents_off': [1, 2, 3, 4, 5, 6, 7]}	0.985	1.0
9	2	N	{'list_of_agents_off': [1, 2, 3, 4, 5]}	0.986	0.98
10	2	N	{'list_of_agents_off': [1, 2, 3]}	0.995	1.0
11	2	N	{'list_of_agents_off': [1]}	0.998	1.0
12	2	O	{'ratio': 0.1, 'interval': 10}	0.97	1.0
13	2	O	{'ratio': 0.2, 'interval': 5}	0.972	0.95
14	2	O	{'ratio': 0.4, 'interval': 5}	0.97	0.98
15	2	O	{'ratio': 0.6, 'interval': 5}	0.981	0.99
16	2	O	{'ratio': 0.8, 'interval': 5}	0.987	0.99
17	2	W	{'ratio': 0.1, 'interval': 10}	0.968	0.99
18	2	W	{'ratio': 0.2, 'interval': 5}	0.964	0.97
19	2	W	{'ratio': 0.4, 'interval': 5}	0.973	0.99
20	2	W	{'ratio': 0.6, 'interval': 5}	0.974	0.98
21	2	W	{'ratio': 0.8, 'interval': 5}	0.985	0.99
22	5	BF	nd.	1.0	1.0
23	5	BFC	nd.	0.929	0.99
24	5	BFO	{'ratio': 0.1, 'interval': 10}	0.926	0.99
25	5	BFO	{'ratio': 0.2, 'interval': 5}	0.924	0.97
26	5	BFO	{'ratio': 0.4, 'interval': 5}	0.926	0.94
27	5	BFO	{'ratio': 0.6, 'interval': 5}	0.94	0.94
28	5	BFO	{'ratio': 0.8, 'interval': 5}	0.963	0.97
29	5	C	nd.	0.92	0.97
30	5	N	{'list_of_agents_off': [1, 2, 3, 4, 5, 6, 7]}	0.951	0.99
31	5	N	{'list_of_agents_off': [1, 2, 3, 4, 5]}	0.967	0.98
32	5	N	{'list_of_agents_off': [1, 2, 3]}	0.977	0.97
33	5	N	{'list_of_agents_off': [1]}	0.993	1.0
34	5	O	{'ratio': 0.1, 'interval': 10}	0.919	0.98
35	5	O	{'ratio': 0.2, 'interval': 5}	0.927	0.93
36	5	O	{'ratio': 0.4, 'interval': 5}	0.934	0.95
37	5	O	{'ratio': 0.6, 'interval': 5}	0.939	0.95
38	5	O	{'ratio': 0.8, 'interval': 5}	0.956	0.96
39	5	W	{'ratio': 0.1, 'interval': 10}	0.931	0.97
40	5	W	{'ratio': 0.2, 'interval': 5}	0.923	0.95
41	5	W	{'ratio': 0.4, 'interval': 5}	0.931	0.93
42	5	W	{'ratio': 0.6, 'interval': 5}	0.946	0.96
43	5	W	{'ratio': 0.8, 'interval': 5}	0.954	0.95
44	10	BF	nd.	1.0	1.0
45	10	BFC	nd.	0.898	0.99
46	10	BFO	{'ratio': 0.1, 'interval': 10}	0.882	0.97
47	10	BFO	{'ratio': 0.2, 'interval': 5}	0.884	0.98
48	10	BFO	{'ratio': 0.4, 'interval': 5}	0.868	0.93
49	10	BFO	{'ratio': 0.6, 'interval': 5}	0.876	0.93
50	10	BFO	{'ratio': 0.8, 'interval': 5}	0.915	0.95
51	10	C	nd.	0.861	0.97
52	10	N	{'list_of_agents_off': [1, 2, 3, 4, 5, 6, 7]}	0.89	0.97
53	10	N	{'list_of_agents_off': [1, 2, 3, 4, 5]}	0.911	0.93

	Parametry ataku			Miary efek.	
Badanie	n_M	atak	parametry ataku	E	$E^{(n)}$
54	10	N	{'list_of_agents_off': [1, 2, 3]}	0.938	0.96
55	10	N	{'list_of_agents_off': [1]}	0.978	0.97
56	10	O	{'ratio': 0.1, 'interval': 10}	0.849	0.94
57	10	O	{'ratio': 0.2, 'interval': 5}	0.851	0.94
58	10	O	{'ratio': 0.4, 'interval': 5}	0.85	0.92
59	10	O	{'ratio': 0.6, 'interval': 5}	0.859	0.87
60	10	O	{'ratio': 0.8, 'interval': 5}	0.919	0.93
61	10	W	{'ratio': 0.1, 'interval': 10}	0.895	0.96
62	10	W	{'ratio': 0.2, 'interval': 5}	0.882	0.98
63	10	W	{'ratio': 0.4, 'interval': 5}	0.872	0.92
64	10	W	{'ratio': 0.6, 'interval': 5}	0.868	0.88
65	10	W	{'ratio': 0.8, 'interval': 5}	0.904	0.91
66	15	BF	nd.	1.0	1.0
67	15	BFC	nd.	0.925	1.0
68	15	BFO	{'ratio': 0.1, 'interval': 10}	0.91	1.0
69	15	BFO	{'ratio': 0.2, 'interval': 5}	0.895	1.0
70	15	BFO	{'ratio': 0.4, 'interval': 5}	0.868	1.0
71	15	BFO	{'ratio': 0.6, 'interval': 5}	0.81	0.89
72	15	BFO	{'ratio': 0.8, 'interval': 5}	0.855	0.86
73	15	C	nd.	0.819	0.93
74	15	N	{'list_of_agents_off': [1, 2, 3, 4, 5, 6, 7]}	0.801	0.93
75	15	N	{'list_of_agents_off': [1, 2, 3, 4, 5]}	0.797	0.9
76	15	N	{'list_of_agents_off': [1, 2, 3]}	0.85	0.9
77	15	N	{'list_of_agents_off': [1]}	0.94	0.95
78	15	O	{'ratio': 0.1, 'interval': 10}	0.806	0.92
79	15	O	{'ratio': 0.2, 'interval': 5}	0.81	0.96
80	15	O	{'ratio': 0.4, 'interval': 5}	0.789	0.88
81	15	O	{'ratio': 0.6, 'interval': 5}	0.779	0.81
82	15	O	{'ratio': 0.8, 'interval': 5}	0.865	0.89
83	15	W	{'ratio': 0.1, 'interval': 10}	0.916	1.0
84	15	W	{'ratio': 0.2, 'interval': 5}	0.904	1.0
85	15	W	{'ratio': 0.4, 'interval': 5}	0.853	0.98
86	15	W	{'ratio': 0.6, 'interval': 5}	0.783	0.85
87	15	W	{'ratio': 0.8, 'interval': 5}	0.848	0.85

W tabeli 16 kolorem czerwonym zaznaczono badania, w których zastosowano atak, który dla danej liczby złośliwych agentów spowodował osiągnięcie w środowisku najniższej efektywności środowiska (E) lub efektywności chwilowej środowiska $n = 100$ ($E^{(n)}$). Warto zwrócić uwagę, że tylko w przypadku istnienia piętnastu złośliwych agentów, największy spadek efektywności w obrębie całego badania (efektywność środowiska), jak i efektywności mierzonej jedynie w końcowej fazie badania (efektywność chwilowa) wystąpił dla tego samego ataku. W pozostałych przypadkach atakujący powinni zastosować inny atak (gdy istnieje dwóch złośliwych agentów) lub przynajmniej inne parametry ataku (gdy istnieje 5 lub 10 złośliwych

agentów) w zależności od ich celu (efekt podczas całej symulacji czy efekt długookresowy). W większości przypadków skuteczny okazał się atak oscylacji zachowania, przy czym parametry tego ataku potrzebne do najbardziej znacznego spadku efektywności środowiska, były różne w zależności od parametrów środowiska (liczby agentów przeprowadzających atak). Atak wyrocznia także był dość efektywny, ale ustępował atakowi oscylacji zachowania prawie we wszystkich przypadkach (wyjątkiem było tylko środowisko złożone z dwóch złośliwych agentów), i to pomimo tego, że jest znacznie bardziej skomplikowany niż pozostałe ataki i znacznie trudniejszy do przeprowadzenia w rzeczywistym środowisku. Wraz ze wzrostem liczby agentów złośliwych stosujących najbardziej efektywny atak (spośród badanych), spadała zarówno średnia efektywność środowiska, jak i efektywność chwilowa środowiska zmierzona na koniec badania. Generalnie należy zauważyć, że system RefTRM, jest dość skuteczny, ponieważ mimo znaczącej przewagi agentów złośliwych (15 agentów w stosunku do 5 agentów rzetelnych), środowisko jest w stanie funkcjonować – spadek efektywności jest wyraźny, ale mimo wszystko efektywność jest znacznie wyższa niż w przypadku braku zastosowania systemu TRM i stosowania najbardziej efektywnego ataku (atak stały, wtedy efektywność wynosi: $E \approx \frac{n_B}{n_B+n_M}$, na mocy twierdzenia 5.1.2.1. i uwagi 5.1.2.2.).

6.3.6. Badanie wpływu parametrów środowiska

Wyniki tego badania (wpływu parametrów środowiska na uzyskane wyniki efektywności systemu TRM) nie będą szczegółowo prezentowane, ale wskazują, że wielkość środowiska (liczba agentów) nie ma wpływu na charakterystykę działania środowiska, czy efektywność poszczególnych ataków. Zwiększenie liczby agentów wpływa jedynie na konieczność przeprowadzenia większej liczby interakcji, aby poszczególne agenty osiągnęły taki sam poziom zaufania do innych agentów (ze względu na większą liczbę możliwych par agentów usługodawca-usługobiorca). Z tego względu, przy wzroście liczby agentów w środowisku pojedyncza interakcja w mniejszym stopniu wpływa na wartości miar globalnego średniego zaufania akcyjnego (i w konsekwencji całkowitego), ale nie na globalne średnie zaufanie rekomendacyjne (z uwagi na to, że i tak w dalszym ciągu rekomendacje są wydawane przez wszystkie agenty). Z uwagi na to, że system RefTRM zakłada ocenę zaufania do wszystkich agentów w środowisku, liczba obliczanych wartości zaufania rośnie potęgowo ze wzrostem liczby agentów w systemie, co ogranicza jego zastosowanie w przypadku dużych (powyżej wielu tysięcy agentów) środowisk.

6.3.7. Tworzenie i badanie ataku dopasowanego

Najbardziej interesujące właściwości systemu RefTRM, które wynikają z jego opisu w punkcie 4.2.9. oraz z opisu środowiska są następujące:

- zaufanie akcyjne do agenta zostanie zwiększone o stałą wartość α^A (do maksymalnej wartości równej 1), o ile jakość usługi przez niego dostarczonej nie będzie mniejsza niż pewien próg będący parametrem systemu: $o \geq h^A$ (na mocy punktu: 4.2.9.8.);
- zaufanie rekomendacyjne do agenta zostanie zwiększone o stałą wartość α^R (do maksymalnej wartości równej 1), o ile różnica pomiędzy jakością dostarczonej usługi i wartością rekomendacji dostarczonej przez tego agenta nie będzie większa niż pewien próg będący parametrem systemu: $\left| r_{a_i \to a_j : a_p}^{c_1 ; m_{l-1}} - o_l \right| \leq h^R$ (na mocy punktu: 4.2.9.7);
- wartość jakości usług oraz rekomendacji może być dowolną wartością z przedziału $\langle 0,1 \rangle$, (na mocy punktów 4.2.9 i 6.2.1).

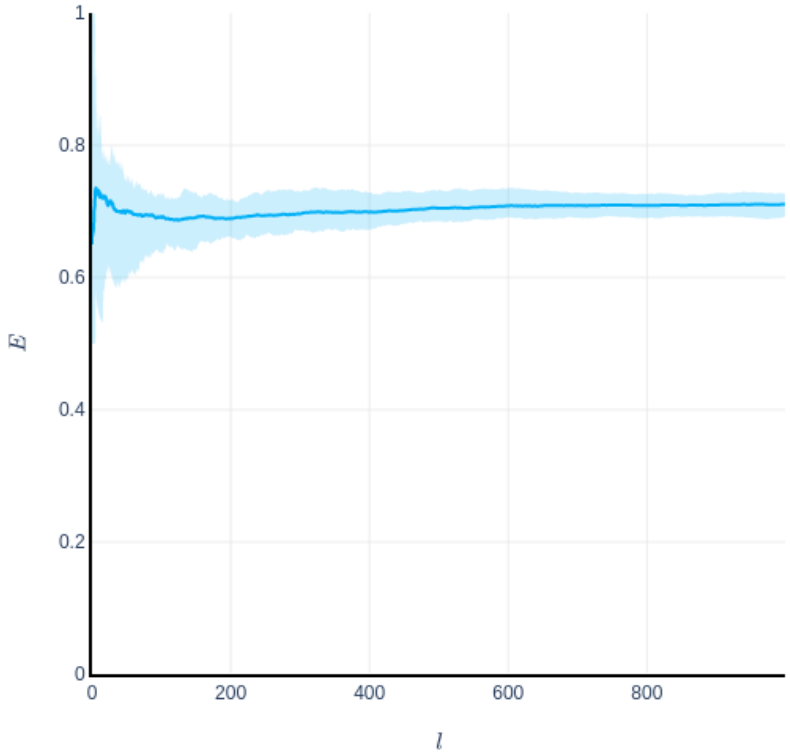
W związku z tym, wydaje się być celowe stworzenie ataku, w którym złośliwe agenty będą:

- świadczyć usługi o jakości $\mathrm{o} = h^A$, wtedy zawsze taka usługa zostanie określona jako rzetelna (nawet gdy $h^A \ll 1$, czyli będzie niższa od maksymalnej jakości usług w środowisku);
- wydawać rekomendacje dotyczące rzetelnych agentów o wartości $r = \max(1 - h^R; 0)$. przy założeniu, że rzetelne agenty zawsze świadczą usługi z maksymalną możliwą jakością, wtedy taka rekomendacja, pomimo że zaniżona, będzie uznana za rzetelną i spowoduje wzrost zaufania rekomendacyjnego do agenta ją wydającego;
- wydawać rekomendacje dotyczące złośliwych agentów o wartości $r = \min(1; h^A + h^R)$, taka rekomendacja, pomimo że zawyżona, będzie uznana za rzetelną i spowoduje wzrost zaufania rekomendacyjnego do agenta ją wydającego.

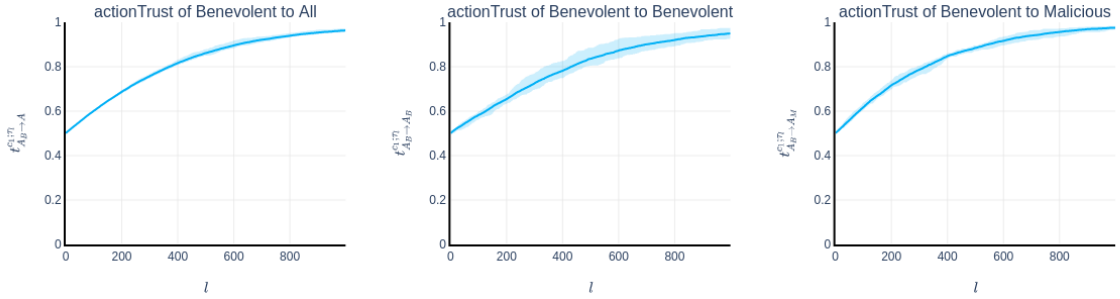
Dla domyślnych wartości parametrów systemu RefTRM, agenty powinny więc świadczyć usługi z jakością $o = 0.5$ oraz wydawać rekomendacje o wartości $r = 0.4$ dotyczące rzetelnych agentów oraz o wartości $r = 1$ dotyczące agentów złośliwych. Atak ten można traktować jako pewien wariant ataku BFCc z dodatkowymi parametrami (jakość usług oraz rekomendacje nie przyjmują wartości skrajnych tak jak w podstawowym wariancie tego ataku badanym wcześniej)[70].

[70] W przypadku ataków C, BF, BFC, a nawet O, BFO, W i N można wprowadzić dodatkowe parametry dotyczące jakości usług, czy wartości rekomendacji i wtedy badać odporność systemu zgodnie z propozycjami badań podanymi w punkcie 6.3.5. Nie zostało to jednak dokonane z powodu, że ataki te w literaturze, rzadko przewidują uwzględnienie takich parametrów, innych niż te które zostały ujęte w opisach tych ataków.

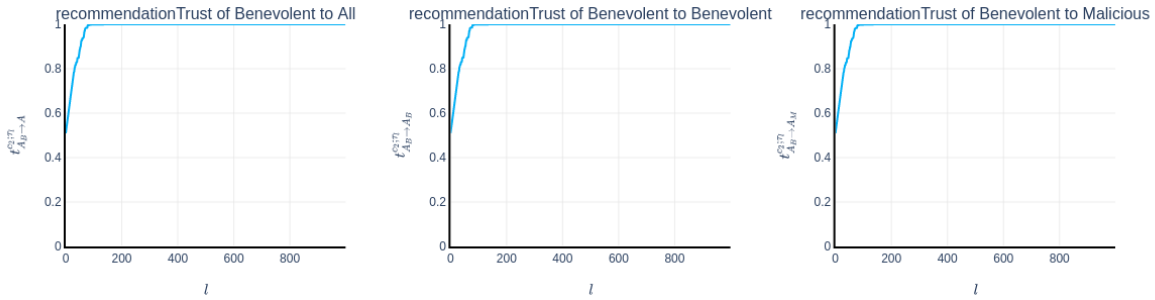
Rysunki 49–53 oraz tabela 17 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak dopasowany (BFC z dodatkowymi parametrami).



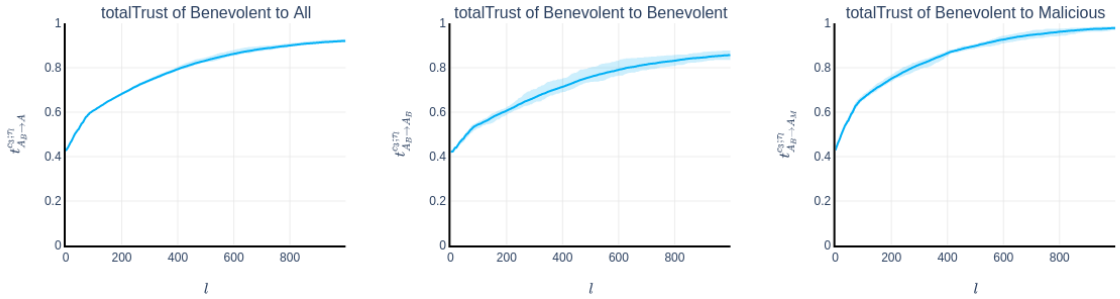
Rysunek 49 Efektywność systemu RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)



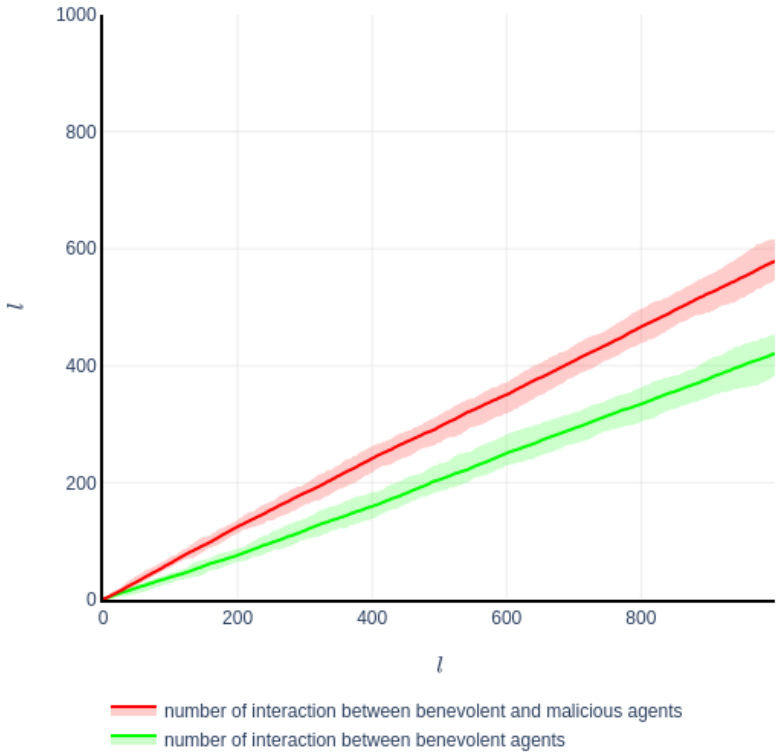
Rysunek 50 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)



Rysunek 51 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)



Rysunek 52 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)



Rysunek 53 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)

Warto zauważyć, że atak ten jest najskuteczniejszy spośród wszystkich do tej pory zaprezentowanych. Świadczy o tym nie tylko znacząco niższa wartość efektywności środowiska, ale także i to, że nie wzrasta ona wraz ze wzrostem liczby interakcji (ma prawie stałą wartość), a dodatkowo rozbieżność pomiędzy efektywnością środowiska, a efektywnością chwilową n=100 pod koniec symulacji, jest niewielka. Ze względu na oczernianie rzetelnych agentów oraz wychwalanie złośliwych agentów, złośliwe agenty uzyskują wyższe wartości zaufania całkowitego agentów rzetelnych niż agenty rzetelne.

Bardzo korzystną cechą tego ataku jest to, że sposób działania złośliwych agentów nie zależy od liczby agentów rzetelnych ani złośliwych, zależy tylko od używanego systemu TRM i jego parametrów (które są znane atakującym) oraz niektórych aspektów środowiska.

Tabela 17 Wyniki badania środowiska z systemem RefTRM w trakcie ataku dopasowanego (BFC z dodatkowymi parametrami)

Symbol	średnia	min	max
E	0.711	0.692	0.727
$E^{(n)}$	0.721	0.68	0.765
$t_{A_B \to A}^{c_1;m_l}$	0.963	0.956	0.973
$t_{A_B \to A_B}^{c_1;m_l}$	0.95	0.927	0.974
$t_{A_B \to A_M}^{c_1;m_l}$	0.976	0.967	0.983
$t_{A_B \to A}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A_B}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A_M}^{c_2;m_l}$	1.0	1.0	1.0
$t_{A_B \to A}^{c_3;m_l}$	0.921	0.914	0.929
$t_{A_B \to A_B}^{c_3;m_l}$	0.857	0.836	0.878
$t_{A_B \to A_M}^{c_3;m_l}$	0.979	0.97	0.986
$l_I^{A_B,A_B}$	421.0	383.0	454.0
$l_I^{A_B,A_M}$	579.0	546.0	617.0

Stworzenie ataku dopasowanego jest nie zawsze możliwe w odniesieniu do każdego systemu TRM, ale powyższy przykład pokazuje, że przy jasnym zdefiniowaniu środowiska oraz systemu (za pomocą przedstawionych modeli) może być łatwiej przeprowadzić taką analizę, a w konsekwencji może być możliwe stworzenie ataku dopasowanego. Natomiast wydaje się, że w głównej mierze to brak jasnej definicji uniemożliwia stworzenie takiego ataku.

Analizując wartości globalnego średniego zaufania (akcyjnego, rekomendacyjnego i całkowitego) agentów rzetelnych do agentów złośliwych, łatwo zauważyć, że agenty złośliwe uzyskują prawie maksymalne wartości poszczególnych typów zaufania. Może to sugerować, że możliwe jest wykorzystanie wysokiego zaufania do jeszcze większego obniżenia efektywności środowiska (świadcząc od czasu do czasu usługę z jakością $o = 0$, zamiast $o = 0.5$). Odpowiednio dobierając częstotliwość tego dodatkowego obniżenia jakości, prawdopodobnie możliwe byłoby cykliczne odzyskiwanie maksymalnych wartości zaufania przez złośliwe agenty (byłby to wtedy pewien wariant ataku BFOc z dodatkowymi parametrami). Dodatkową przesłanką mogącą wskazywać, że taki atak byłby jeszcze bardziej skuteczny jest analiza wyników badań w punkcie 6.3.5, zawartych w tabeli 15. Wyniki te sugerują, że atak BFOc jest bardziej skuteczny niż atak BFCc (obydwa jednak były

przeprowadzane bez uwzględnienia dodatkowych parametrów). Łatwo zweryfikować twierdząco tę hipotezę przeprowadzając nowe badanie, ale autor odstąpił od prezentowania jego wyników, z uwagi na to, że odkrycia podobnego ataku o podobnej skuteczności, dokonano w oparciu o metodę MEAEM (niezależnie od tworzenia ataku dopasowanego).

6.3.8. Badanie z uogólnionym atakiem – metodą MEAEM

Badania środowiska i systemu RefTRM z użyciem metody MEAEM zostały wykonane w dwóch wariantach:

- z wykorzystaniem analizy możliwych zachowań atakujących odnośnie świadczenia usługi i wydawania rekomendacji (wariant 1), oraz
- z wykorzystaniem analizy możliwych zachowań atakujących jedynie odnośnie świadczenia usługi (wariant 2).

Na potrzeby metody MEAEM, na bazie analizy parametrów systemu RefTRM i środowiska zdecydowano się na użycie następujących zbiorów wartości składających się na zestaw decyzji:

- $R_{A_M} = \{0; 0.4; 0.6; 1.0\}$ – zbiór wartości rekomendacji wydawanych przez złośliwe agenty, które mogą zostać dostarczone dla rzetelnych agentów o złośliwych agentach, $\left|R_{A_M}\right| = k_{R_{A_M}} = 4$;
- $R_{A_B} = \{0; 0.4; 0.6; 1.0\}$ – zbiór wartości rekomendacji wydawanych przez złośliwe agenty, które mogą zostać dostarczone dla rzetelnych agentów o rzetelnych agentach, $\left|R_{A_B}\right| = k_{R_{A_B}} = 4$;
- $Q_M = \{0; 0.5; 1.0\}$ – zbiór jakości usług, które mogą zostać wyświadczone przez agentów złośliwych dla agentów rzetelnych, $|Q_M| = k_{Q_M} = 3$.

Oczywiście zbiory te można znacząco rozszerzyć, wtedy nie ma potrzeby opierać się na analizie funkcjonowania systemu TRM, jednak wydłuża to znacznie czas badań (co ma szczególnie duże znaczenie w przypadku badań według wariantu 1).

W systemie RefTRM występują trzy konteksty zaufania (akcyjne, rekomendacyjne i całkowite), przy czym zaufanie całkowite jest obliczane na podstawie zaufania akcyjnego, zaufania rekomendacyjnego i rekomendacji. Jak zauważono w trakcie opisu metody MEAEM w podrozdziale 5.4., stosowane miary zaufania w wyrażeniu na wartość funkcji zysku atakujących powinny być niezależne. Z tego względu wydaje się celowe aby przy konstrukcji wyrażenia na wartość funkcji zysku brać pod uwagę zaufanie akcyjne i rekomendacyjne lub

tylko zaufanie całkowite. Zdecydowano się na tę pierwszą możliwość z uwagi na to, że wartości zaufania akcyjnego i rekomendacyjnego są łatwiejsze w interpretacji, a dodatkowo wpływ na ten wybór miała techniczna architektura narzędzia, która powodowała, że aby dokonać aktualizacji wartości zaufania całkowitego dla wszystkich agentów, konieczne byłoby przeprowadzenie dodatkowych obliczeń, co w połączeniu ze znaczną liczbą analizowanych przypadków w odniesieniu do zestawu decyzji, spowodowałoby dodatkowe wydłużenie badań.

W celu dokonania określenia wartości współczynników istotności, dokonano symulacji kilku pierwszych interakcji dla różnych zestawów decyzji i zaobserwowano, że efektywność chwilowa $n = 100$ zmienia się w związku z jedną interakcją w zakresie od 0 do 0.01, globalne średnie zaufanie akcyjne w zakresie od -0.004 do 0.0022, a globalne średnie zaufanie rekomendacyjne w zakresie od -0.025 do 0.01. Zdecydowano się na przypisanie wyższej wartości współczynnika istotności efektywności, tak aby agenty złośliwe miały za cel minimalizację tej wartości oraz na dodatkowe zmniejszenie współczynnika istotności zaufania rekomendacyjnego z uwagi na to, że ma ono znacznie mniejszy niż zaufanie akcyjne wpływ na zaufanie całkowite, które jest wykorzystywane do podejmowania decyzji przez agenty o wyborze usługodawcy. Wobec tego przyjęto następujące wartości współczynników istotności[71]: γ_E =1; $\gamma_{t^{c_1}} = 1$; $\gamma_{t^{c_2}} = 0.1$; $\gamma_{t^{c_3}} = 0$ i zastosowano następujące wyrażenie pozwalające obliczyć wartość funkcji zysku atakujących podczas podejmowania decyzji dotyczącej dostarczania rekomendacji i jakości świadczonej usługi:

$$
\begin{aligned}
f_g{}^{r+a} = \%_{A_M}^{m_l} \Big(& -\Delta E_{/A_M}^{(n):m_l} - \Delta t_{A_B \to A_B/A_M}^{c_1;m_l} - 0.1\Delta t_{A_B \to A_B/A_M}^{c_2;m_l} + \Delta t_{A_B \to A_M/A_M}^{c_1;m_l} \\
& + 0.1\Delta t_{A_B \to A_M/A_M}^{c_2;m_l} \Big) \\
& + \%_{A_B}^{m_l} \Big(-\Delta E_{/A_B}^{(n):m_l} - \Delta t_{A_B \to A_B/A_B}^{c_1;m_l} - 0.1\Delta t_{A_B \to A_B/A_B}^{c_2;m_l} + \Delta t_{A_B \to A_M/A_B}^{c_1;m_l} \\
& + 0.1\Delta t_{A_B \to A_M/A_B}^{c_2;m_l} \Big)
\end{aligned}
$$

W przypadku obliczania wartości funkcji zysku atakujących podczas podejmowania decyzji dotyczącej jedynie jakości świadczonej usługi, zastosowano następujące wyrażenie:

$$
f_g{}^{a} = -\Delta E_{/A_M}^{(n):m_l} - \Delta t_{A_B \to A_B/A_M}^{c_1;m_l} - 0.1\Delta t_{A_B \to A_B/A_M}^{c_2;m_l} + \Delta t_{A_B \to A_M/A_M}^{c_1;m_l} + 0.1\Delta t_{A_B \to A_M/A_M}^{c_2;m_l}
$$

Warto zwrócić uwagę, że w przypadku wariantu 2 analiza zgodnie z metodą MEAEM dotyczyła jedynie tych interakcji, w których wybrany został złośliwy agent jako usługodawca, podczas gdy w przypadku wariantu 1, analiza była wykonywana podczas każdej interakcji.

[71] Rozszerzone badania (wyników nie zaprezentowano w niniejszej pracy), w których użyto innych współczynników istotności ujawniły, że dobór współczynników ma niewielki wpływ na wyniki działania metody MEAEM. Kilkukrotne zwiększenie względnej wartości współczynników względem pozostałych współczynników nie wpłynęło w zauważalnym stopniu na wyniki metody.

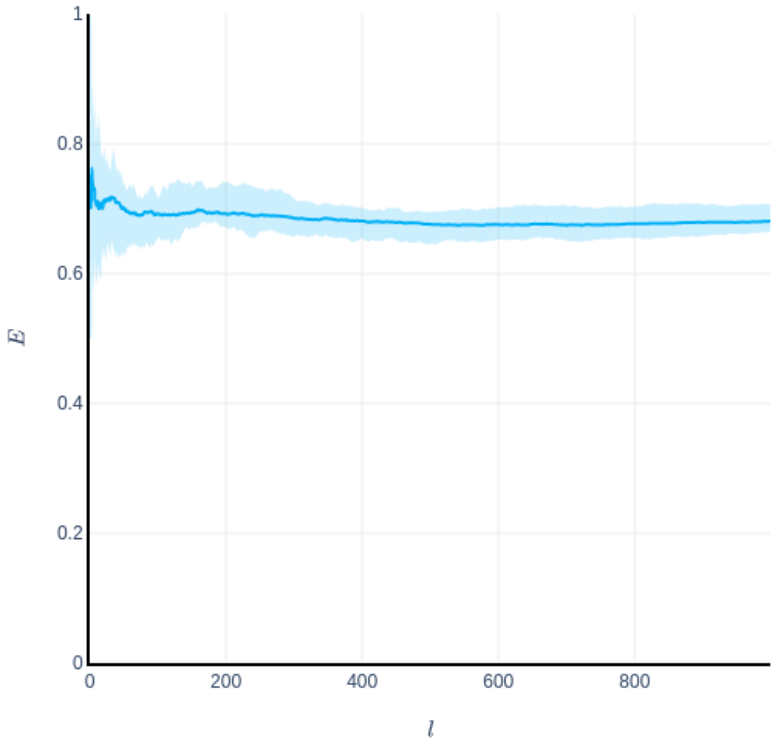
Dodatkowo w przypadku wariantu 1 konieczne było zasymulowanie zarówno wybrania złośliwego jak i rzetelnego agenta w każdym rozważanym przypadku (dla każdego zestawu decyzji). Liczba przypadków, co oczywiste, w ramach pojedynczej interakcji także byłaby znacząco wyższa w przypadku pierwszego wariantu. Wszystko to sprawiło, że o ile wykonanie pojedynczego badania metodą MEAEM w wariancie 2 trwało kilka minut, to w przypadku wariantu 1 czas wydłużał się do kilku godzin[72]. Rysunki 54–58 oraz tabela 18 prezentują wyniki otrzymane z badania środowiska z funkcjonującym systemem RefTRM, w przypadku gdy złośliwe agenty wykonują atak metodą MEAEM. Zaprezentowane wyniki zostały otrzymane za pomocą wariantu 2.

Wyniki otrzymane za pomocą wariantu 1 były nieznacznie gorsze z perspektywy atakujących (tzn. efektywność środowiska wynosiła około 0.72, co było wynikiem wręcz nieznacznie gorszym niż w przypadku zastosowania ataku dopasowanego). Ten, na pierwszy rzut oka, nieco zaskakujący wynik da się jednak łatwo wyjaśnić. W wyrażeniu na wartość funkcji zysku atakujących brane były pod uwagę zagregowane miary zaufania w postaci średniego globalnego zaufania agentów rzetelnych odpowiednio do agentów rzetelnych i złośliwych oraz zagregowane prawdopodobieństwa wyboru jako usługodawcy agenta rzetelnego lub złośliwego. Mogło to spowodować, że mimo wyboru optymalnego zestawu decyzji w przypadku gdy zostałby wybrany jako usługodawca pewien agent złośliwy, to został wybrany inny agent, do którego rzetelne agenty miały inne wartości zaufania. Jak wykazano już na etapie opisu metody MEAEM rozważenie indywidualnego zachowania agentów złośliwych byłoby trudne z uwagi na potęgowy wzrost liczby analizowanych przypadków. Z tego względu, mimo że następowały próby manipulacji zarówno rekomendacjami, jak i jakością usług, to nie zawsze były one trafne. Warto zaznaczyć, że w trakcie badań zgodnie z metodą MEAEM wedle wariantu pierwszego, złośliwe agenty starały się wykonywać działania podobne, do tych zaprezentowanych w ataku wyrocznia, tzn. jednocześnie obniżać rekomendacje do złośliwych agentów i w przypadku gdy złośliwy agent został wybrany wtedy świadczyć usługi o minimalnej jakości. Jednak działania te nie odniosły zamierzonego skutku tzn. nie doprowadziły do większego zmanipulowania agentów rzetelnych. Dlatego prostszy atak zgodnie z wariantem drugim metody MEAEM mógł okazać się skuteczniejszy.

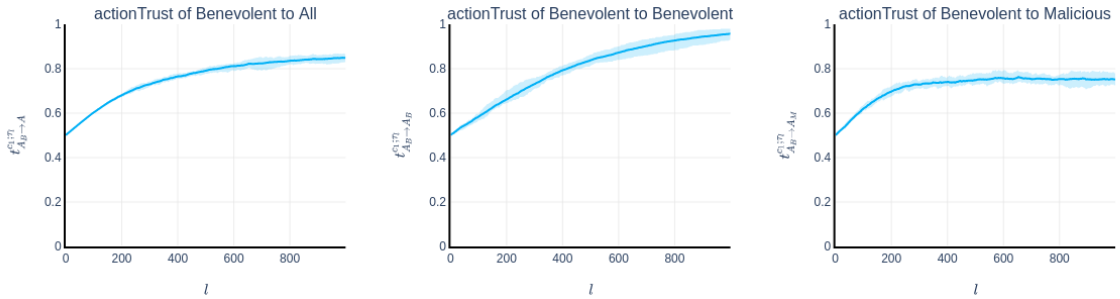
W przypadku badań z użyciem metody MEAEM według wariantu drugiego używano rekomendacji złośliwych agentów o innych złośliwych agentach o wartości 1.0, a o rzetelnych agentach o wartości 0.4, czyli takich jak w przypadku ataku dopasowanego. Użycie innych

[72] Szczegóły techniczne dotyczące sprzętu na jakim zostały przeprowadzone badania, zawarto w załączniku 5.

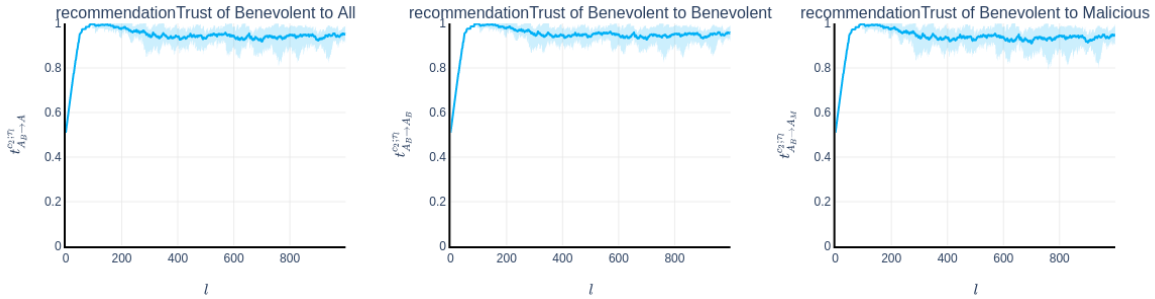
wartości rekomendacji np. w obu przypadkach o wartości 1.0, miało niewielki wpływ na efektywność (znacząco mniejszy niż wahania wartości efektywności uzyskane w różnych przebiegach badania).



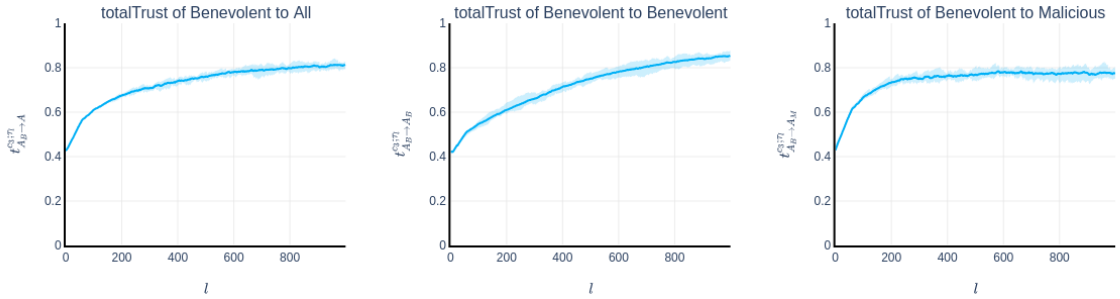
Rysunek 54 Efektywność systemu RefTRM w trakcie ataku metodą MEAEM



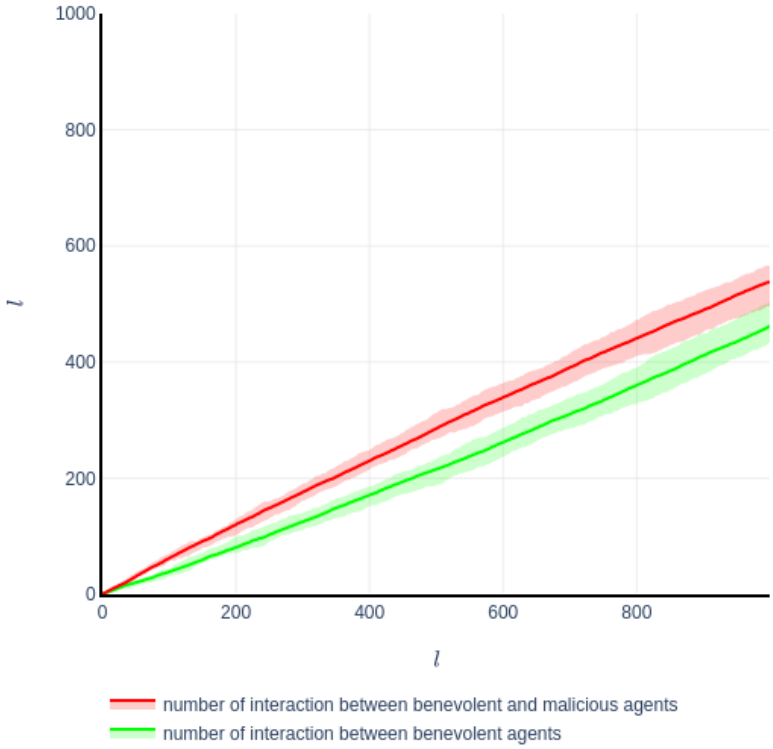
Rysunek 55 Globalne średnie zaufanie akcyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku metodą MEAEM



Rysunek 56 Globalne średnie zaufanie rekomendacyjne agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku metodą MEAEM



Rysunek 57 Globalne średnie zaufanie całkowite agentów rzetelnych, odpowiednio do wszystkich agentów, agentów rzetelnych oraz agentów złośliwych dla systemu RefTRM w trakcie ataku metodą MEAEM



Rysunek 58 Liczba interakcji pomiędzy agentami rzetelnymi oraz agentami rzetelnymi i złośliwymi w przypadku funkcjonowania systemu RefTRM w trakcie ataku metodą MEAEM

Tabela 18 Wyniki badania środowiska z systemem RefTRM w trakcie ataku metodą MEAEM

Symbol	**średnia**	**min**	**max**
E	0.681	0.666	0.708
$E^{(n)}$	0.699	0.6	0.745
$t_{A_B \to A}^{c_1;m_l}$	0.849	0.831	0.865
$t_{A_B \to A_B}^{c_1;m_l}$	0.958	0.931	0.98
$t_{A_B \to A_M}^{c_1;m_l}$	0.751	0.726	0.779
$t_{A_B \to A}^{c_2;m_l}$	0.953	0.914	0.993
$t_{A_B \to A_B}^{c_2;m_l}$	0.958	0.926	0.991
$t_{A_B \to A_M}^{c_2;m_l}$	0.948	0.902	0.994
$t_{A_B \to A}^{c_3;m_l}$	0.813	0.794	0.826
$t_{A_B \to A_B}^{c_3;m_l}$	0.854	0.834	0.876

$t_{A_B \to A_M}^{c_3;m_l}$	0.776	0.753	0.796
$l_I^{A_B,A_B}$	462.0	434.0	504.0
$l_I^{A_B,A_M}$	538.0	496.0	566.0

Atak opracowany za pomocą metody MEAEM okazał się skuteczniejszy od wszystkich poprzednich ataków. Nie powinno to być zaskakujące, ponieważ złośliwe agenty dążyły do optymalizacji własnych działań, tak aby z jednej strony cechować się wysoką wartością zaufania, ale z drugiej strony maksymalnie obniżać efektywność środowiska. Na czym polegał w praktyce odkryty atak z pomocą metody MEAEM? W przypadku wariantu drugiego (którego wyniki zostały zaprezentowane) złośliwe agenty stosowały rekomendacje dokładnie tak jak w przypadku ataku dopasowanego (zawyżanie zaufania do agentów złośliwych poprzez wydawanie rekomendacji o wartości 1, oraz zaniżanie zaufania do agentów rzetelnych poprzez wydawanie rekomendacji o tych agentach o wartości 0.4). Agenty złośliwe generalnie świadczyły usługi z jakością taką jak w ataku dopasowanym, czyli o jakości $q = 0.5$, ale w przypadku gdy po świadczeniu usługi, która zostałaby uznana jako rzetelna, wartość średniego zaufania agentów rzetelnych do agentów złośliwych nie wzrosłaby znacząco[73], to agent ten świadczył usługę z najniższą możliwą jakością. W dalszej części badania (po tym jak agenty złośliwe uzyskały wysokie zaufanie akcyjne) każdy z agentów złośliwych średnio co trzecią usługę świadczył o najniższej jakości, a pozostałe dwie o satysfakcjonującej jakości ($q = 0.5$). W związku z tym, agenty po początkowym okresie zdobywania zaufania zaczęły stosować atak oscylacji zachowania z dodatkowymi parametrami (nigdy nie świadczyły usług z maksymalną możliwą jakością).

6.3.9. Podsumowanie i ocena wyników badań

Tabela 19 zawiera zestawienie wyników badań środowiska z systemem RefTRM z domyślnymi parametrami. W tabeli zawarto wyniki miar wiarygodności w przypadku stosowania przez atakujących najskuteczniejszego dobrze znanego ataku z domyślnymi parametrami – czyli ataku oscylacji zachowania; ataku dopasowanego, który został opracowany na podstawie analizy systemu RefTRM oraz ataku zidentyfikowanego za pomocą metody MEAEM. Jak wskazano wcześniej i co potwierdzają uzyskane wyniki, ataki próbujące wykorzystać specyficzne właściwości systemu (tzn. atak dopasowany i atak MEAEM) mogą

[73] Zachodziło to w momencie gdy wartość zaufania akcyjnego agenta będącego usługobiorcą do agenta będącego usługodawcą była bliska maksymalnej wartości, czyli 1, wobec tego nie było możliwe dalsze zwiększanie tego zaufania, co powodowało, że złośliwemu agentowi chwilowo „przestawało opłacać się" świadczenie usługi o jakości, która zostałaby uznana jako rzetelna.

okazać się znacznie bardziej skuteczne, zmniejszając efektywność środowiska używającego dany system i wskazywać na ograniczoną wiarygodność danego systemu TRM.

Tabela 19 Zestawienie wyników badania środowiska z systemem RefTRM w przypadku niektórych ataków

Symbol	Miara	**Atak O**	**Atak dopasowany (BFC*)**	**Atak metodą MEAEM**
E	efektywność środowiska	0.849	0.711	0.681
$E^{(n)}$	efektywność chwilowa n	0.911	0.721	0.699
$t_{A_B \to A}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do wszystkich agentów	0.558	0.963	0.849
$t_{A_B \to A_B}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do agentów rzetelnych	0.994	0.95	0.958
$t_{A_B \to A_M}^{c_1;m_l}$	globalne średnie zaufanie akcyjne agentów rzetelnych do agentów złośliwych	0.165	0.976	0.751
$t_{A_B \to A}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do wszystkich agentów	0.985	1.0	0.953
$t_{A_B \to A_B}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów rzetelnych	0.969	1.0	0.958
$t_{A_B \to A_M}^{c_2;m_l}$	globalne średnie zaufanie rekomendacyjne agentów rzetelnych do agentów złośliwych	1.0	1.0	0.948
$t_{A_B \to A}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do wszystkich agentów	0.543	0.921	0.813
$t_{A_B \to A_B}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do agentów rzetelnych	0.907	0.857	0.854
$t_{A_B \to A_M}^{c_3;m_l}$	globalne średnie zaufanie całkowite agentów rzetelnych do agentów złośliwych	0.215	0.979	0.776
$l_I^{A_B,A_B}$	liczba interakcji agentów rzetelnych z agentami rzetelnymi	747.8	421.0	462.0
$l_I^{A_B,A_M}$	liczba interakcji agentów rzetelnych z agentami złośliwymi	252.2	579.0	538.0

Jak pokazuje przykład badań, standardowe ataki opisywane w literaturze (nawet te wysublimowane) mogą okazać się znacznie mniej skuteczne niż atak dopasowany do danego systemu TRM lub otrzymany za pomocą metody MEAEM. Z tego względu metoda MEAEM wydaje się być obiecująca, gdyż nie wymaga ona dokładnego przeanalizowania systemu TRM (tak jak w przypadku tworzenia ataku dopasowanego). Na podstawie wykonanych badań można stwierdzić, że najbardziej efektywnym atakiem na system RefTRM z domyślnymi parametrami jest atak przeprowadzony za pomocą metody MEAEM[74]. W związku z tym możliwe jest

[74] Oczywiście nie można wykluczyć, że istnieje takie działanie atakujących, które umożliwiłoby większe zakłócenie działania środowiska, gdyż metoda MEAEM wykorzystywała heurystykę. Niemniej jednak na podstawie analizy wyników badań w przypadku stosowania różnych ataków, trudno odnaleźć symptomy wskazujące, że takie działanie mogłoby istnieć.

wyznaczenie zysku efektywności i zysku absolutnego efektywności w przypadku stosowania w środowisku systemu RefTRM, zgodnie z definicjami 5.1.2.2. i 5.1.2.3. Zestawienie wartości efektywności osiągniętej w środowisku podczas funkcjonowania systemu RefTRM i bez takiego systemu, w czasie gdy atakujący wykonują atak stały (najbardziej efektywny atak w przypadku braku systemu TRM) i atak MEAEM (najbardziej efektywny atak w przypadku funkcjonowania systemu RefTRM z domyślnymi parametrami), zawarto w tabeli 20.

Tabela 20 Wyznaczanie miar zysku efektywności

		Wartości średnie ze wszystkich przebiegów badania		
System TRM	**Atak**	**Wartość efektywności**	**Średnia jakość usług świadczonych przez złośliwe agenty**	**Liczba interakcji agentów rzetelnych z agentami złośliwymi**
brak	**stały**	$E_{0,} = 0.479$	0	Około $\frac{10}{19}$ liczby wszystkich interakcji
brak	**MEAEM**[75]	$E_0 = 0.687$	Przyjęto tak jak w ataku MEAEM w systemie RefTRM: 0.407	Przyjęto tak jak w ataku stałym – około $\frac{10}{19}$ liczby wszystkich interakcji
RefTRM	**MEAEM**	$E_{+TRM} = 0.681$	0.407[76]	538

Atak MEAEM jest atakiem tworzonym dynamicznie w odniesieniu do danego systemu RefTRM. W przypadku środowiska bez funkcjonującego systemu metoda ta albo nie będzie możliwa do przeprowadzenia (jeżeli założymy, że warunkiem koniecznym jest uzależnienie funkcji zysku od pewnej wartości zaufania, która w takim przypadku nie występuje), albo wygeneruje atak stały (wartość funkcji zysku będzie zależała jedynie od efektywności chwilowej, wobec czego najbardziej opłacalne dla atakujących będzie zawsze świadczenie usługi z minimalną jakością). W związku z tym, przyjęto, że rozważone zostanie takie samo zachowanie agentów w środowisku bez systemu TRM, jak w przypadku gdy funkcjonuje system RefTRM, tzn. że agenty świadczą usługi z pewną średnią jakością (ta wartość została wzięta z analizy dokonanej metodą MEAEM w środowisku z systemem RefTRM, czyli wynosi 0.407) podczas działania środowiska, ale liczba interakcji ze złośliwymi agentami została

[75] Wartość została obliczona, a nie uzyskana w badaniu, wyjaśnienie znajduje się w dalszej części punktu.

[76] Wartość można uzyskać z badań lub przeprowadzić następujące rozumowanie: agenty rzetelne świadczą usługi o jakości 1.0 i wyświadczyły średnio 462 usługi, wobec tego agenty złośliwe podczas średnio 538 interakcji wyświadczyły usługi o sumarycznej jakości $(681 - 462 = 219)$, wobec tego średnia jakość usług wyniosła: 0.407

wyznaczona na podstawie specyfiki tego środowiska (i wynosi $\frac{10}{19}$ wszystkich interakcji czyli około: 526 interakcji). W pozostałych interakcjach agenty rzetelne będą świadczyć usługi z maksymalną jakością, co da w rezultacie wartość efektywności równą[77]: $E_0 = \frac{10*0,407+9*1}{19} = 0.687$.

Na tej podstawie wyznaczmy:

- zysk efektywności, który wynosi: $G = E_{+TRM} - E_0 = 0.681 - 0.687 = -0.006$, oraz
- zysk absolutny efektywności: $G_A = E_{+TRM} - E_{0'} = 0.681 - 0.479 = 0.202$

Jak widać, w przypadku gdyby agenty zachowywały się dokładnie w ten sam sposób, niezależnie od tego czy w środowisku działa system RefTRM, czy też nie, to efektywność środowiska byłaby na zbliżonym poziomie, a zysk efektywności byłby bliski zeru, a wręcz ujemny (co oznacza, że gdyby atakujący stosowali ten atak, to z punktu widzenia efektywności systemu lepiej, żeby system RefTRM wtedy nie był używany przez agenty), co jest interesującą obserwacją. Nie oznacza to jednak, że zastosowanie systemu RefTRM jest zupełnie bezzasadne, gdyż w przypadku jego braku istnieje bardziej efektywny atak na takie środowisko (atak stały).

Przeprowadzone badania pozwoliły zidentyfikować interesujące właściwości systemu RefTRM i wskazać jego podatności, między innymi to, że stosując atak dopasowany lub atak wygenerowany przez metodę MEAEM, możliwe jest istotne zaburzenie działania środowiska.

[77] Możliwe byłoby też przeprowadzenia badania symulacyjnego, w którym złośliwe agenty w przypadku braku systemu TRM zachowywałyby się tak jak w przypadku badania metodą MEAEM z systemem RefTRM, tzn. stosowałaby atak oscylacji zachowania z dodatkowymi parametrami (świadcząc usługi z jakością maksymalną 0.5 i minimalną 0), ale takie postępowanie nie jest niezbędne, bo na bazie analizy teoretycznej też możliwe jest wyznaczenie efektywności w przypadku tego środowiska i zachowania agentów.

7. PODSUMOWANIE I WNIOSKI

W rozprawie zostały opracowane modele środowiska, systemu TRM i ataku, które pozwalają na ustandaryzowanie ich opisów, a w konsekwencji na badania różnych systemów TRM. Praca prezentuje propozycję autora w zakresie metodyki oceny wiarygodności systemów TRM, która być może przyczyni się do uporządkowania zagadnień związanych z oceną odporności systemów TRM na ataki. Wskazane zostały elementy badań mechanizmów zarządzania zaufaniem, które powinny być wykonane w celu oceny ich odporności na ataki. Mimo tego, że praca wskazuje pewne przesłanki uzasadniające przyjęcie określonych sposobów badań i miar, to nie przedstawia w sposób ścisły dowodów ich adekwatności (z uwagi na naturalne ograniczenia: ogólność modelu systemu TRM i ogólność modelu środowiska). Trudno definitywnie stwierdzić, że zaproponowany przez autora zestaw badań jest wystarczający w odniesieniu do wszystkich możliwych systemów TRM, stanowi jednakże pewien zestaw wytycznych, które pozwalają na przeprowadzanie oceny wiarygodności dużej grupy systemów TRM w sposób metodyczny.

Pewnym ograniczeniem metodyki, jak również metody symulacyjnej, za pomocą której zostały wykonane badania w niniejszej rozprawie, jest fakt, że siłą rzeczy przy badaniu danego systemu TRM, funkcjonującego w określonym środowisku, nie mogą być uwzględnione wszystkie możliwe aspekty (np. bardzo wiele charakterystyk żądań agentów, czy różne topologie relacji agentów, itd.). Jednakże wyniki badań prowadzonych w różnych warunkach, nie pokazują znaczących rozbieżności, co pozwala wysnuć hipotezę, że przy zmienionej charakterystyce żądań agentów (których oczywiście nie da się określić a priori), jak i przy zmienionej wielkości środowiska (liczbie agentów wchodzących w interakcje), uzyskiwane wyniki badań nie będą różnić się znacząco. Takie podejście uprawnia do wyciągania uogólnionych wniosków co do efektywności danego systemu TRM w określonym środowisku, bez potrzeby posiadania wiedzy o wszystkich warunkach w danym środowisku. Niemniej, jednym z zaleceń proponowanej metodyki jest przeprowadzanie badań, symulując na możliwie zbliżonym poziomie dane środowisko (np. przeprowadzając badania w środowisku o zbliżonej liczbie agentów i pewnej ogólnej charakterystyce żądań). Wydaje się, że takie podejście jest skuteczne w uzyskaniu wyników na potrzeby praktycznego porównania systemów TRM lub wskazania ich głównych braków. Należy jednak mieć świadomość, że efektywność danego systemu TRM określona w rzeczywistym środowisku może się w pewnej mierze różnić od tej uzyskanej w trakcie badań symulacyjnych.

Z jednej strony rozprawa definiuje sposób opisu systemu TRM, aby była możliwa weryfikacja jego wiarygodności, z drugiej strony definiuje sposób opisu ataku. Praca wskazuje znane ataki przeciwko TRM, ale także prezentuje metodę identyfikacji nowych ataków, dopasowanych do konkretnego systemu TRM (której wyniki muszą być każdorazowo zweryfikowane w celu oceny praktycznej realizowalności ataku). Zawiera też propozycję narzędzia i przykładowe badania przeprowadzone z jego użyciem.

Przedstawione wyniki badań dla przykładowego środowiska i systemu TRM, pokazują przydatność przedstawionych modeli, metodyki oceny wiarygodności oraz działanie w praktyce metody identyfikacji nowych ataków – MEAEM, jak również adekwatność wytycznych dotyczących identyfikacji nowego ataku na podstawie analizy teoretycznej konkretnego systemu TRM.

7.1. PODSUMOWANIE ORYGINALNYCH WYNIKÓW AUTORA

Głównymi osiągnięcia rozprawy jest opracowanie modelu środowiska, systemu TRM i ataku oraz zdefiniowanie miar wiarygodności systemów TRM. Przyjęte uogólnienia i modele pozwalają na porównywanie wiarygodności różnych systemów, zarówno w badaniach symulacyjnych jak i w rzeczywistych środowiskach. Kolejnym istotnym osiągnięciem autora jest opracowanie architektury narzędzia TRM-RET, która umożliwia niezależne definiowanie systemów TRM oraz ataków na te systemy. Dzięki temu badanie nowego systemu TRM wymaga jedynie zaimplementowania samego systemu, natomiast nie ma potrzeby implementacji poszczególnych znanych ataków oddzielnie dla każdego systemu. Oczywiście wciąż istnieje możliwość zaimplementowania nowego ataku, który będzie mógł zostać wykorzystany w odniesieniu do wszystkich zaimplementowanych systemów (bez konieczności ich aktualizacji). Ostatnim istotnym osiągnięciem pracy jest stworzenie metody heurystycznej MEAEM, pozwalającej na odkrycie nowego, potencjalnie najbardziej efektywnego, ataku na dany system TRM. Istotną wartością pracy jest też pokazanie, że mimo tego, że przykładowy system okazał się dość odporny na znane standardowe ataki, to udało się go skutecznie zmanipulować stosując atak opracowany przez metodę MEAEM lub atak dopasowany.

7.2. PUBLIKACJE I WYSTĄPIENIA AUTORA ZWIĄZANE Z TEMATYKĄ ROZPRAWY

Autor publikował artykuły i wygłaszał wystąpienia konferencyjne związane z tematyką systemów zarządzania zaufaniem i reputacją. Prace autora w tym zakresie można podzielić na dwie główne grupy:

- ściśle związane z tematyką rozprawy, prace dotyczące oceny wiarygodności systemów TRM oraz metodyki badań ich efektywności, w tym koncepcje ataków na systemy TRM lub metod je wykrywających;
- prace prezentujące propozycje nowych systemów TRM lub specyficznych zastosowań zaufania np. do oceny rzetelności informacji, w tym artykuły, w których prezentowane metody oceny zaufania są jedynie fragmentem całej publikacji.

Do pierwszej grupy należą publikacje dotyczące:

- prekursorów narzędzia TRM-RET stworzonych przez autora: [93], [109];
- opracowania modelu systemów oraz sposobów oceny odporności na ataki i propozycji ocen wiarygodności systemów TRM: [13], [66], [106], [107], które jednakże zostały znacząco rozwinięte w trakcie prac nad rozprawą;
- propozycji ataku wyrocznia: [45], [85];
- prezentacji ogólnej koncepcji metody wykrywania ataków na systemy TRM, będącej wstępem do metody MEAEM: [83];
- zestawienia i porównania ataków na systemy TRM: [79];
- propozycji metody służącej do wykrywania kooperacyjnych ataków na systemy TRM: [71].

Warto podkreślić, że niniejsza praca jest także poniekąd odpowiedzią na uwagi recenzentów dotyczące artykułów [66], [83], wskazujące, że przedstawione metryki i narzędzie to istotny element tworzonej metodyki, ale wciąż niekompletny, gdyż nie zostały w ramach tych artykułów określone zasady postępowania (zalecenia) dotyczące definiowania konkretnego systemu oraz prowadzenia eksperymentów dla różnych ataków. Niniejsza praca stanowi, zdaniem autora, usystematyzowanie wiedzy dotyczącej metodycznego przeprowadzania analiz odporności systemów TRM na ataki.

Wśród artykułów należących do drugiej grupy należy wymienić:

- publikacje dotyczące zastosowania metod oceny zaufania do informacji o podatnościach i eksploitach, w celu zagregowania, skorelowania, wybrania, rozszerzenia i oceny rzetelności informacji pochodzących z różnych źródeł: [43], [44];

- publikację prezentującą ideę oceny zaufania do aktywów teleinformatycznych (konkretnego oprogramowania, urządzeń, itd.) i ryzyka związanego z ich użytkowaniem, na bazie informacji historycznych o zidentyfikowanych podatnościach oraz dojrzałości procesu zarządzania podatnościami przez danego producenta: [51], a także publikację dotyczącą wykorzystania tego podejścia do szacowania ryzyka konkretnych podmiotów na bazie identyfikacji wykorzystywanych aktywów teleinformatycznych: [110];
- propozycję metody umożliwiającej integrację i konsolidację informacji z różnych platform e-handlu, które wykorzystują systemy zarządzania zaufaniem i reputacją w celu zapewnienia bezpieczeństwa użytkowników: [111].

Niektóre z powyższych prac zostały zaprezentowane przez autora na konferencjach międzynarodowych, m.in.: IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID); Annual Computer Security Applications Conference (ACSAC); European Interdisciplinary Cybersecurity Conference (EICC); Conference on Cryptography and Security Systems (CSS); Central European Conference on Cryptology (CECC); IEEE-SPIE Joint Symposium on Photonics, Web Engineering, Electronics for Astronomy and High Energy Physics Experiments oraz konferencji: Krajowe Sympozjum Telekomunikacji i Teleinformatyki (KSTIT).

7.3. PERSPEKTYWY KONTYNUACJI BADAŃ

Jak wskazano w rozprawie, interesującą perspektywą kontynuacji badań jest sprawdzenie możliwości wykorzystania wielu systemów TRM jednocześnie w jednym środowisku i zweryfikowanie jak to przełoży się na efektywność poszczególnych agentów i całego środowiska. Naturalną kontynuacją prac byłoby zbadanie, zgodnie z metodyką, kolejnych systemów TRM i porównanie ich właściwości, w tym uzyskanych w trakcie badań wartości miar wiarygodności dla różnych systemów funkcjonujących w różnych środowiskach. Interesującym zagadnieniem jest także rozwinięcie modelu systemu TRM oraz narzędzia, tak aby istniała możliwość uwzględnienia niezerowego czasu trwania interakcji. Zagadnieniem, które wydaje się najbardziej interesujące dla autora rozprawy jest funkcjonowanie wielu niezależnych grup złośliwych agentów w jednym środowisku i badania wpływu ich działań na efektywność środowiska i funkcjonowanie systemu TRM.

BIBLIOGRAFIA

[1] Y. L. Sun, Z. Han, W. Yu, and K. J. R. Liu, "Attacks on trust evaluation in distributed networks," *Inf. Sci. Syst. 2006 40th Annu. Conf.*, pp. 1461–1466, 2006.

[2] W. Michael, *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2009.

[3] D. J. Bianco, "The Pyramid of Pain," *Enterprise Detection & Response, dostęp: 30.06.2022*, 2013, http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html.

[4] L. Rasmusson and S. Jansson, "Simulated social control for secure Internet commerce," *Proc. 1996 Work. New Secur. Paradig. - NSPW '96*, pp. 18–25, 1996.

[5] M. A. Tunia, "Nowy model adaptacyjny usługi niezaprzeczalności wykorzystującej twarde i miękkie mechanizmy bezpieczeństwa," 2018.

[6] M. Szymczak, "Słownik języka polskiego PWN." Wydawnictwa naukowe PWN, 1995.

[7] "Słownik języka polskiego on-line," *dostęp: 30.06.2022*, https://sjp.pwn.pl/.

[8] A. Wierzbicki, *Trust and fairness in open, distributed systems*, vol. 298. Springer, 2010.

[9] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A survey of attack and defense techniques for reputation systems," *ACM Comput. Surv.*, vol. 42, no. 1, pp. 1–31, 2009.

[10] Y. Sun and Y. Liu, "Security of online reputation systems: The evolution of attacks and defenses," *IEEE Signal Process. Mag.*, vol. 29, no. 2, pp. 87–97, 2012.

[11] T. Zahariadis, P. Trakadas, H. Leligou, P. Karkazis, and S. Voliotis, "Implementing a Trust-Aware Routing Protocol in Wireless Sensor Nodes," *2010 Dev. E-systems Eng.*, pp. 47–52, 2010.

[12] M. Blaze, J. Feigenbaum, J. Ioannidis, and A. D. Keromytis, "The Role of Trust Management in Distributed Systems Security," in *Secure Internet Programming: Security Issues for Mobile and Distributed Objects*, J. Vitek and C. D. Jensen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 185–210.

[13] M. Janiszewski, "Metamodel Systemów Zarządzania Zaufaniem i Reputacją," *Przegląd Telekomun. i Wiadomości Telekomun.*, vol. LXXXVI, no. 8-9/2017, pp. 803–808, 2017.

[14] E. Ayday, H. Lee, and F. Fekri, "An iterative algorithm for trust and reputation management," *Inf. Theory, 2009. ISIT 2009. IEEE Int. Symp.*, pp. 2051–2055, 2009.

[15] E. Ayday and F. Fekri, "Iterative Trust and Reputation Management Using Belief Propagation," *Dependable Secur. Comput. IEEE Trans.*, vol. 9, no. 3, pp. 375–386, 2012.

[16] J. I. Khan and S. S. Shaikh, "A multi-scenario reputation estimation framework and its resilience study against various forms of attacks," *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. WI 2007*, pp. 676–682, 2007.

[17] S. Liu, H. Yu, C. Miao, and A. C. Kot, “A fuzzy logic based reputation model against unfair ratings,” *Proc. 2013 Int. Conf. Auton. agents multi-agent Syst.*, pp. 821–828, 2013.

[18] Y. Liu, Y. L. Sun, S. Liu, and A. C. Kot, “Securing Online Reputation Systems Through Trust Modeling and Temporal Analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 936–948, 2013.

[19] “Strona internetowa aplikacji Yanosik,” *dostęp: 30.06.2022*, www.yanosik.pl.

[20] Van Kranenburg R., E. Anzelmo, A. Bassi, D. Caprio, S. Dodson, and M. Ratto, “The internet of things. A critique of ambient technology and the all-seeing network of RFID,” *Netw. Notebooks*, 2008.

[21] S. Dyrda, L. Wisłowski, and J. Zawiła-Niedźwiecki, “Teleinformatyczne technologie przyszłości w zarządzaniu,” in *Metody badania i modele rozwoju organizacji*, 2012, pp. 93–108.

[22] D. Jain, P. V. Krishna, and V. Saritha, “A Study on Internet of Things based Applications,” *CoRR*, 2012.

[23] M. Janiszewski, “Bezpieczeństwo w Internecie Rzeczy,” in *Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej. Technologie Informacyjne, tom 21, Wydanie na temat: Architektura Internetu Przyszłości*, 2013, pp. 121–137.

[24] D. Kozlov, J. Veijalainen, and Y. Ali, “Security and Privacy Threats in IoT Architectures,” *Proc. 7th Int. Conf. Body Area Networks*, pp. 256–262, 2012.

[25] L. Atzori, A. Iera, G. Morabito, and M. Nitti, “The social internet of things (SIoT) - When social networks meet the internet of things: Concept, architecture and network characterization,” *Comput. Networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

[26] W. Leister and T. Schulz, “Ideas for a Trust Indicator in the Internet of Things,” in *SMART 2012 : The First International Conference on Smart Systems, Devices and Technologies Ideas*, 2015, no. December, pp. 31–34.

[27] J. Chen, Z. Tian, X. Cui, L. Yin, and X. Wang, “Trust architecture and reputation evaluation for internet of things,” *J. Ambient Intell. Humaniz. Comput.*, vol. 0, no. 0, pp. 1–9, 2018.

[28] D. Chen, G. Chang, D. Sun, J. Li, J. Jia, and X. Wang, “TRM-IoT: A trust management model based on fuzzy reputation for internet of things,” *Comput. Sci. Inf. Syst.*, vol. 8, no. 4, pp. 1207–1228, 2011.

[29] N. Djedjig, D. Tandjaoui, I. Romdhani, and F. Medjek, “Trust Management in Internet of Things,” 2012.

[30] A. Riahi, Y. Challal, E. Natalizio, Z. Chtourou, and A. Bouabdallah, “A Systemic

Approach for IoT Security," *2013 IEEE Int. Conf. Distrib. Comput. Sens. Syst.*, pp. 351–355, 2013.

[31] U. E. Tahta, S. Sen, and A. B. Can, "GenTrust: A genetic trust management model for peer-to-peer systems," *Appl. Soft Comput. J.*, vol. 34, pp. 693–704, 2015.

[32] A. B. Can and B. Bhargava, "SORT: A self-organizing trust model for peer-to-peer systems," *IEEE Trans. Dependable Secur. Comput.*, vol. 10, no. 1, pp. 14–27, 2013.

[33] X. Fan, L. Liu, M. Li, and Z. Su, "EigenTrustp++: Attack resilient trust management," *Collab. Comput. Networking, Appl. Work. (CollaborateCom), 2012 8th Int. Conf.*, pp. 416–425, 2012.

[34] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust algorithm for reputation management in P2P networks," *Proc. twelfth Int. Conf. World Wide Web - WWW '03*, p. 640, 2003.

[35] H. Marzi and M. Li, "An enhanced bio-inspired trust and reputation model for wireless sensor network," *Procedia Comput. Sci.*, vol. 19, pp. 1159–1166, 2013.

[36] X. He, X. Gui, and W. Wei, "A heider-theory based reputation framework for WSN," *Proc. - 10th IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2008*, pp. 635–640, 2008.

[37] S. Sudharani, V. George Samuel Raj, and S. Earnest Paul, "A Lightweight Trust System with Provisioning for Detecting Malicious Node in Clustered Wireless Sensor Networks," vol. 21, no. 7, pp. 335–341, 2015.

[38] H. Wang and Y. Zhang, "AraTRM: Attack resistible ant-based trust and reputation model," *Proc. - 2014 IEEE Int. Conf. Comput. Inf. Technol. CIT 2014*, pp. 652–657, 2014.

[39] P. Ebinger and N. Bißmeyer, "TEREC: Trust evaluation and reputation exchange for cooperative intrusion detection in MANETs," *Proc. 7th Annu. Commun. Networks Serv. Res. Conf. CNSR 2009*, pp. 378–385, 2009.

[40] A. Banerjee, S. Neogy, and C. Chowdhury, "Reputation based trust management system for MANET," *Proc. - 2012 3rd Int. Conf. Emerg. Appl. Inf. Technol. EAIT 2012*, pp. 376–381, 2012.

[41] B. Yang, R. Yamamoto, and Y. Tanaka, "Dempster-Shafer evidence theory based trust management strategy against cooperative black hole attacks and gray hole attacks in MANETs," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. V, pp. 223–232, 2014.

[42] X. Chen, J. H. Cho, and S. Zhu, "GlobalTrust: An attack-resilient reputation system for tactical networks," *2014 11th Annu. IEEE Int. Conf. Sensing, Commun. Networking,*

SECON 2014, pp. 275–283, 2014.

[43] M. Janiszewski, A. Felkner, P. Lewandowski, M. Rytel, and H. Romanowski, "Automatic Actionable Information Processing and Trust Management towards Safer Internet of Things," *Sensors 2021, Vol. 21*, vol. 21, no. 13, pp. 43–59, Jun. 2021.

[44] M. Rytel, A. Felkner, and M. Janiszewski, "Towards a Safer Internet of Things—A Survey of IoT Vulnerability Data Sources," *Sensors 2020, Vol. 20*, vol. 20, no. 21, pp. 59–69, Oct. 2020.

[45] M. Janiszewski, "The Oracle a New Intelligent Cooperative Strategy of Attacks on Trust and Reputation Systems," *Ann. UMCS, Inform.*, vol. 14, no. 2, pp. 86–101, 2014.

[46] T. Kavitha and D. Sridharan, "Security Vulnerabilities In Wireless Sensor Networks : A Survey," *J. Inf. Assur. Secur.*, vol. 5, no. 2010, pp. 31–44, 2010.

[47] M. Janiszewski, "Bezpieczeństwo w Internecie Rzeczy," *Przegląd Telekomun. i Wiadomości Telekomun.*, no. 8-9/2013, pp. 727–734, 2013.

[48] S. I. Ahamed, M. M. Haque, and N. Talukder, "A formal context specific Trust model (FTM) for multimedia and ubiquitous computing environment," *Telecommun. Syst.*, vol. 44, no. 3–4, pp. 221–240, 2010.

[49] X. Wang, W. Cheng, P. Mohapatra, and T. Abdelzaher, "Enabling reputation and trust in privacy-preserving mobile sensing," *IEEE Trans. Mob. Comput.*, vol. 13, no. 12, pp. 2777–2790, 2014.

[50] X. Wang, L. Liu, and J. Su, "RLM: A general model for trust representation and aggregation," *IEEE Trans. Serv. Comput.*, vol. 5, no. 1, pp. 131–143, 2012.

[51] M. Janiszewski, A. Felkner, and J. Olszak, "Trust and Risk Assessment Model of Popular Software Based on Known Vulnerabilities," *Int. J. Electron. Telecommun.*, vol. 63, no. 3, pp. 329–336, 2017.

[52] T. Muller, J. Zhang, and Y. Liu, "A language for trust modelling," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10003 LNAI, pp. 1–12, 2016.

[53] L. Mui, "Computational Models of Trust and Reputation : Agents , Evolutionary Games , and Social Networks," *Soc. Networks*, p. 139, 2002.

[54] M. Tavakolifard and K. C. Almeroth, "A taxonomy to express open challenges in trust and reputation systems," *J. Commun.*, vol. 7, no. 7 SUPPL.7, pp. 538–551, 2012.

[55] A. Bravetti and P. Padilla, "An optimal strategy to solve the Prisoner's Dilemma," *Sci. Rep.*, vol. 8, no. 1, p. 1948, 2018.

[56] F. Gómez Mármol and G. Martínez Pérez, "Towards pre-standardization of trust and

reputation models for distributed and heterogeneous systems," *Comput. Stand. Interfaces*, vol. 32, no. 4, pp. 185–196, 2010.

[57] Y. Yu, K. Li, W. Zhou, and P. Li, "Trust mechanisms in wireless sensor networks: Attack analysis and countermeasures," *J. Netw. Comput. Appl.*, vol. 35, no. 3, pp. 867–880, 2012.

[58] F. G. Marmol and G. M. Perez, "State of the art in trust and reputation models in P2P networks," in *Handbook of Peer-to-Peer Networking*, 2010, pp. 761–784.

[59] J. Sabater and C. Sierra, "Review on computational trust and reputation models," *Artif. Intell. Rev.*, vol. 24, no. 1, pp. 33–60, 2005.

[60] M. Momani and S. Challa, "Survey of trust models in different network domains," vol. 1, no. 3, 2010.

[61] Z. Noorian and M. Ulieru, "The state of the art in trust and reputation systems: A framework for comparison," *J. Theor. Appl. Electron. Commer. Res.*, vol. 5, no. 2, pp. 97–117, 2010.

[62] B. Yu and M. P. Singh, "Detecting deception in reputation management," *Proc. Second Int. Jt. Conf. Auton. agents multiagent Syst. - AAMAS '03*, p. 73, 2003.

[63] X. Jiang and L. Ye, "Reputation-based trust model and anti-attack mechanism in P2P networks," *NSWCTC 2010 - 2nd Int. Conf. Networks Secur. Wirel. Commun. Trust. Comput.*, vol. 1, pp. 498–501, 2010.

[64] R. Magdich, H. Jemal, and M. Ben Ayed, "A resilient Trust Management framework towards trust related attacks in the Social Internet of Things," *Comput. Commun.*, vol. 191, no. May 2021, pp. 92–107, 2022.

[65] S. A. Ghasempouri and B. Tork Ladani, "Modeling trust and reputation systems in hostile environments," *Futur. Gener. Comput. Syst.*, vol. 99, no. May, pp. 571–592, 2019.

[66] M. Janiszewski, "Towards an evaluation model of trust and reputation management systems," *Int. J. Electron. Telecommun.*, vol. 63, no. 4, pp. 411–416, 2017.

[67] F. G. Mármol and G. M. Pérez, "TRMSim-WSN, trust and reputation models simulator for wireless sensor networks," *IEEE Int. Conf. Commun.*, 2009.

[68] M. Sievers, "Modeling Trust and Reputation in Multiagent Systems," in *Handbook of Model-Based Systems Engineering*, A. M. Madni, N. Augustine, and M. Sievers, Eds. Cham: Springer International Publishing, 2022, pp. 1–36.

[69] Y. Sun, Z. Han, and K. J. Liu, "Defense of trust management vulnerabilities in distributed networks," *IEEE Commun. Mag.*, vol. 46, no. 2, pp. 112–119, 2008.

[70] L. Zhang, S. Jiang, J. Zhang, and W. K. Ng, “Robustness of trust models and combinations for handling unfair ratings,” *IFIP Adv. Inf. Commun. Technol.*, vol. 374 AICT, pp. 36–51, 2012.

[71] M. Janiszewski, “SimGroupTest - nowa metoda detekcji kooperacyjnych ataków złośliwych węzłów przeciwko systemom zarządzania zaufaniem,” *Przegląd Telekomun. - Wiadomości Telekomun.*, vol. LXXXVIII, no. 8-9/2015, pp. 822–829, 2015.

[72] A. Srinivasan, J. Teitelbaum, H. Liang, J. Wu, and M. Cardei, “Reputation and Trust-based Systems for Ad Hoc and Sensor Networks,” *Algorithms Protoc. Wireless, Mob. Ad Hoc Networks*, pp. 375–403, 2008.

[73] P. B. B. Velloso, R. P. P. Laufer, O.-C. M. B. Duarte, and G. Pujolle, “A Trust Model Robust to Slander Attacks in Ad Hoc Networks,” *2008 Proc. 17th Int. Conf. Comput. Commun. Networks*, vol. 6, p. 121, 2008.

[74] F. G. Mármol and G. M. Pérez, “Security threats scenarios in trust and reputation models for distributed systems,” *Comput. Secur.*, vol. 28, no. 7, pp. 545–556, 2009.

[75] M. Srivatsa, L. Xiong, and L. Liu, “TrustGuard: countering vulnerabilities in reputation management for decentralized overlay networks,” *Proc. 14th Int. Conf. World Wide Web*, pp. 422–431, 2005.

[76] Y. Yang, Q. Feng, Y. Sun, and Y. Dai, “Reputation trap: An powerful attack on reputation system of file sharing p2p environment,” *Proc. 4th Int. Conf. Secur. Priv. Commun. Networks*, pp. 1766–1780, 2008.

[77] H. K. Jnanamurthy and S. Singh, “Detection and filtering of collaborative malicious users in reputation system using quality repository approach,” *Proc. 2013 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2013*, pp. 466–471, 2013.

[78] R. Kerr and R. Cohen, “Smart Cheaters Do Prosper : Defeating Trust and Reputation Systems,” *Proc. 8th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS 2009)*, pp. 993–1000, 2009.

[79] M. Janiszewski, “Klasyfikacja i ocena efektywności ataków na systemy zarządzania zaufaniem,” in *Prace Seminarium Naukowego Instytutu Telekomunikacji Politechniki Warszawskiej. Tom 1*, A. Jakubiak, Ed. Oficyna Wydawnicza PW, 2015, pp. 101–115.

[80] S. Babar, P. Mahalle, A. Stango, N. Prasad, and R. Prasad, “Proposed security model and threat taxonomy for the Internet of Things (IoT),” *Commun. Comput. Inf. Sci.*, vol. 89 CCIS, pp. 420–429, 2010.

[81] A. A. Pirzada and C. McDonald, “Establishing trust in pure ad-hoc networks,” *Proc. 27th Australas. Conf. Comput. Sci.*, vol. 26, no. c, pp. 47–54, 2004.

[82] D. G. Padmavathi and M. D. Shanmugapriya, “A Survey of Attacks, Security Mechanisms and Challenges in Wireless Sensor Networks,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 4, no. 1, pp. 1–9, 2009.

[83] M. Janiszewski, “MEAEM – Metoda Identyfikacji i Oceny Najbardziej Efektywnego Ataku Przeciwko Systemom Zarządzania Zaufaniem i Reputacją,” *Przegląd Telekomun. i Wiadomości Telekomun.*, no. 8-9/2016, pp. 852–858, 2016.

[84] N. K. Saini, V. K. Sihag, and R. C. Yadav, “A reactive approach for detection of collusion attacks in P2P trust and reputation systems,” *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 312–317, 2014.

[85] M. Janiszewski, “Atak wyrocznia jako przykład strategii ataku opartego na kooperacji złośliwych węzłów przeciwko systemom zarządzania zaufaniem,” *Przegląd Telekomun. i Wiadomości Telekomun.*, no. 8-9/2014, pp. 1006–1013, 2014.

[86] N. K. Saini, “Identifying Collusion Attack in P2P Trust and Reputation Systems,” pp. 36–41, 2014.

[87] G. Costagliola, V. Fuccella, and F. A. Pascuccio, “Towards a trust, reputation and recommendation meta model,” *J. Vis. Lang. Comput.*, vol. 25, no. 6, pp. 850–857, 2014.

[88] P.-L. Sun and C.-Y. Ku, “Review of threats on trust and reputation models,” *Ind. Manag. Data Syst.*, vol. 114, no. 3, pp. 472–483, 2014.

[89] A. Jalaly Bidgoly and B. Tork Ladani, “Benchmarking reputation systems: A quantitative verification approach,” *Comput. Human Behav.*, vol. 57, pp. 274–291, 2016.

[90] A. Jøsang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decis. Support Syst.*, vol. 43, no. 2, pp. 618–644, 2007.

[91] T. Zahariadis, H. C. Leligou, S. Voliotis, S. Maniatis, P. Trakadas, and P. Karkazis, “An Energy and Trust-aware Routing Protocol for Large Wireless Sensor Networks,” *9th WSEAS Int. Conf. Appl. Informatics Commun. (AIC ’09)*, pp. 216–224, 2009.

[92] J. Konorski, “Reputacja i zaufanie w systemach teleinformatycznych z podmiotami anonimowymi - podejście dynamiczne,” *Przegląd Telekomun. + Wiadomości Telekomun.*, vol. 8-9/2026, pp. 698–712, 2016.

[93] M. Janiszewski, “TRM-EAT - narzędzie oceny odporności na ataki i efektywności systemów zarządzania zaufaniem,” *Przegląd Telekomun. - Wiadomości Telekomun.*, vol. LXXXVIII, no. 8-9/2015, pp. 813–821, 2015.

[94] F. Gómez Mármol and G. Martínez Pérez, “Trust and reputation models comparison,” *Internet Res.*, vol. 21, no. 2, pp. 138–153, 2011.

[95] K. K. Fullam, T. Klos, G. Muller, J. Sabater-Mir, K. S. Barber, and L. Vercouter, “The

Agent Reputation and Trust (ART) Testbed," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3986 LNCS, pp. 439–442, 2006.

[96] K. K. Fullam *et al.*, "A specification of the Agent Reputation and Trust (ART) testbed: experimentation and competition for trust in agent societies," *Proc. fourth Int. Jt. Conf. Auton. agents multiagent Syst.*, pp. 512–518, 2005.

[97] C. W. Hang, Y. Wang, and M. P. Singh, "An adaptive probabilistic trust model and its evaluation," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, 2008, pp. 1485–1488.

[98] M. Harbers, R. Verbrugge, C. Sierra, and J. Debenham, "The examination of an information based approach to trust," in *P. Noriega, J. Padget, (Red.), International workshop on coordination, organization, institutions and norms (COIN)*, 2008, pp. 101–112.

[99] R. Kerr and R. Cohen, "TREET: The Trust and Reputation Experimentation and Evaluation Testbed," *Electron. Commer. Res.*, vol. 10, no. 3, pp. 271–290, 2010.

[100] A. A. Adamopoulou and A. L. Symeonidis, "A simulation testbed for analyzing trust and reputation mechanisms in unreliable online markets," *Electron. Commer. Res. Appl.*, vol. 13, no. 5, pp. 368–386, 2014.

[101] S. Antipolis, "EStarMom : Extendable Simulator for Trust and Reputation Management in Online Marketplaces," pp. 1–10, 2014.

[102] A. Irissappane and J. Zhang, "A testbed to evaluate the robustness of reputation systems in e-marketplaces," *Proc. 13th Inter- Natl. Conf. Auton. agents multi-agent Syst. AAMAS*, pp. 1629–1630, 2014.

[103] Y. Wang and J. Vassileva, "Trust and reputation model in peer-to-peer networks," *Proc. - 3rd Int. Conf. Peer-to-Peer Comput. P2P 2003*, pp. 150–157, 2003.

[104] J. Głowacka, M. Amanowicz, J. Krygier, and M. Głowacki, "Badanie mechanizmu obrony przed zagrożeniami węzła taktycznej sieci doraźnej," *Przegląd Telekomun. i Wiadomości Telekomun.*, vol. 8-9/2014, pp. 1142–1147, 2014.

[105] J. Sen, "A distributed trust and reputation framework for mobile Ad Hoc networks," *Commun. Comput. Inf. Sci.*, vol. 89 CCIS, pp. 538–547, 2010.

[106] M. Janiszewski, "Metody oceny ataków przeciwko systemom zarządzania zaufaniem i reputacją," in *Prace Seminarium Naukowego Instytutu Telekomunikacji Politechniki Warszawskiej. Tom 2*, A. Jakubiak, Ed. Oficyna Wydawnicza PW, 2016, pp. 88–105.

[107] M. Janiszewski, "Methods for reliability evaluation of trust and reputation systems," in

Proceedings of SPIE - The International Society for Optical Engineering, 2016, vol. 10031, p. 100314B.

[108] Y. L. Sun, Z. Han, W. Yu, and K. J. R. Liu, "Attacks on trust evaluation in distributed networks," in *2006 IEEE Conference on Information Sciences and Systems, CISS 2006 - Proceedings*, 2007.

[109] M. Janiszewski, "TRM-EAT-a New Tool for Reliability Evaluation of Trust and Reputation Management Systems in Mobile Environments," in *Proceedings - 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGRID 2020*, 2020, pp. 718–727.

[110] M. Janiszewski, A. Felkner, and P. Lewandowski, "A novel approach to national-level cyber risk assessment based on vulnerability management and threat intelligence," *J. Telecommun. Inf. Technol.*, no. 2, pp. 5–14, 2019.

[111] M. Janiszewski, "Algorytm konsolidacji ocen zaufania i reputacji w platformach e-handlu," *Przegląd Telekomun. i Wiadomości Telekomun.*, no. 8-9/2016, pp. 934–937, 2016.

[112] Y. L. Y. L. Sun, Z. Z. Han, W. Yu, and K. J. R. K. Liu, "A trust evaluation framework in distributed networks: Vulnerability analysis and defense against attacks," *Proc. IEEE INFOCOM 2006. 25TH IEEE Int. Conf. Comput. Commun.*, vol. 6, no. c, pp. 1–13, 2006.

ZAŁĄCZNIKI

ZAŁĄCZNIK 1 – WYKAZ UŻYWANYCH SKRÓTÓW

Skrót	**Znaczenie**
CSIRT	Computer Security Incident Response Team – zespół reagowania na incydenty komputerowe
C&C	command and control
DB	DataBase – baza danych
DGA	Domain Generation Algorithm
DoS	Denial of Service
IDS	Intrusion Detection System – system wykrywania intruzów
IPS	Intrusion Prevention System – system zapobiegania intruzom
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
IoC	Indicator of Compromise – wskaźnik kompromitacji
IoT	Internet of Things – Internet Rzeczy
JSON	JavaScript Object Notation
ODS	Open Distributed Systems
MA	Malicious Agent – złośliwy agent
MANET	Mobile Ad hoc Networks – mobilne sieci doraźne
MEAEM	Most Effective Attack Evaluation Method – metoda oceny najbardziej efektywnego ataku
P2P	Peer-to-peer
RA	Reliable Agent – rzetelny agent
RMSS	Recommendation Management Subsystem – podsystem zarządzania rekomendacjami
SIEM	Security Information and Event Management
TRM	Trust and Reputation Management – zarządzanie zaufaniem i reputacją
TRM-RET	Trust and Reputation Management – Reliability Evaluation Testbed
TTP	Tactics, Techniques and Procedures – taktyki, techniki i procedury
WSN	Wireless Sensor Networks – sieci bezprzewodowych czujników
Oznaczenia ataków	
B	Bad-mouthing – oczernianie
BF	Bad-mouthing, False-praise – atak oczerniania i wychwalania
BFCc	Bad-mouthing, False-praise, constant with cooperation – atak oczerniania, wychwalania i stały z kooperacją
BFOc	Bad-mouthing, False-praise, on-off with cooperation – atak oczerniania, wychwalania i oscylacji zachowania z kooperacją
C	Constant – atak stały
F	False-praise – wychwalanie
H	wHitewashing – atak kreacji nowej tożsamości
M	Atak MEAEM
N	coNflicting bahaviour - niespójne zachowanie
O	On-off – oscylacja zachowania
S	Sybil – atak kreacji wielu tożsamości
W	Wyrocznia

ZAŁĄCZNIK 2 – WYKAZ OZNACZEŃ

Niniejszy załącznik zawiera wykaz oznaczeń, które zostały zastosowane w pracy.

Oznaczenie	**Wyjaśnienie**	**Zależności lub definicja**	**Miejsce w rozprawie**
A	zbiór agentów	$\|A\| = n$	R4.1.
$a_1, \dots, a_n$	agenty o numerach odpowiednio $1, .., n$	$A = \{a_1, \dots, a_n\}$	R4.1.
n	liczba agentów w środowisku		R4.1.
A_B	zbiór agentów rzetelnych	$A_B \subseteq A$, $\|A_B\| = n_B$	R4.1.
A_M	zbiór agentów złośliwych	$A_M \subseteq A$, $\|A_M\| = n_M$ $A_B \cap A_M = \emptyset$ $A_B \cup A_M = A$ $n_B + n_M = n$	R4.1.
$u_1, \dots, u_m$	usługi o numerach odpowiednio $1, .., m$		R4.1.
U	zbiór usług	$U = \{u_1, \dots, u_l\}$	R4.1.
U_{a_k}	zbiór usług świadczonych przez agenta a_k	$U_{a_k} \subseteq U$ $\bigcup_{i=1}^{n} U_{a_i} = U$	R4.1.
Q	zbiór możliwych jakości usług		R4.1.
q_{min}	minimalna jakość usług w środowisku	$q_{min} \in Q$	R4.1.
q_{max}	maksymalna jakość usług w środowisku	$q_{max} \in Q$	R4.1.
q_0	jakość usługi odpowiadająca brakowi usługi	$q_0 \in Q$	R4.1.
$q_{a_k}^{u_l}$	maksymalna jakość usługi u_l, świadczonej przez agenta a_k	$q_{a_k}^{u_l} \in Q$	R4.1.
A_P	zbiór agentów – usługodawców	$\forall a_k \in A_P: U_{a_k} \neq \emptyset$, $A_P \subseteq A$, $A_P \neq \emptyset$	R4.1.
$A_{P:u_l}$	zbiór agentów – usługodawców usługi u_l	$\forall a_k \in A_{P:u_l}: u_l \in U_{a_k}$	R4.1.
$U_{R:a_k}$	usługi żądane przez agenta a_k	$U_{R:a_k} \subseteq U$	R4.1.
A_R	zbiór agentów – usługobiorców	$\forall a_k \in A_R: U_{R:a_k} \neq \emptyset$, $A_R \subseteq A$ $\bigcup_{i=1}^{n} U_{R:a_i} \subseteq U$	R4.1.
m_l	moment czasowy interakcji o numerze l		R4.1.2.
m_0	czas rozpoczęcia działania środowiska		R4.1.2.
M	zbiór czasów kolejnych żądań (interakcji)	$M = \{m_1, m_2, \dots, m_{l_I}\}$	R4.1.2.
l_I	liczba interakcji w środowisku lub	$l_I \in \mathbb{N}$	R4.1.2.

	numer ostatniej interakcji		
L_I	zbiór numerów wszystkich interakcji	$L_I = \{1,2,\dots,l_I\}$ $\lvert L_I\rvert = l_I$	R4.1.2.
e_l	żądanie	$e_l = (a_i\ u_k, m_l)$ gdzie: $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana usługa, $m_l \in M$ – czas pojawienia się żądania o numerze l	Definicja 4.1.2., R4.1.2.
E	zbiór wszystkich żądań	$E = \{e_1, e_2, \dots, e_{l_I}\}$ $\lvert E\rvert = l_I$	Uwaga 4.1.2.2., R4.1.2.
f_{int} $f_{int}(a_i, u_k, m_l, a_j)$	Funkcja interakcji (funkcja częściowa)	$f_{int}: A \times U \times M \times A \rightarrow Q$ przy czym jeżeli $f_{int}(a_i, u_k, m_l, a_j) = q^l$, to $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana (świadczona) usługa, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę), a $q^l \in Q$ – jakość dostarczonej usługi w ramach l-tej interakcji (wynik tej interakcji).	Definicja 4.1.3.1., R4.1.3.
f_{intE} $f_{intE}(e_l, a_j)$	interakcja w oparciu o żądanie (funkcja częściowa)	$f_{intE}: E \times A \rightarrow Q$ przy czym $e_l \in E$ – żądanie o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę), a $q_l \in Q$ – jakość dostarczonej usługi w ramach tej interakcji (wynik tej interakcji)	Uwaga 4.1.3.1., R4.1.3.
I	zbiór interakcji	Dziedzina funkcji interakcji	Definicja 4.1.3.2., R4.1.3.
i_l	interakcja o numerze l (argument funkcji interakcji o numerze l	krotka: (a_i, u_k, m_l, a_j) przy czym $a_i \in A$ – usługobiorca (agent żądający usługi), $u_x \in U$ – żądana (świadczona) usługa, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę),	Definicja 4.1.3.3., R4.1.3.
Q_{RES}	ciąg wyników interakcji	$Q_{RES}: L_I \rightarrow Q$ gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru Q i odpowiadają kolejnym wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji	Definicja 4.1.3.4., R4.1.3.
q^l	wynik interakcji o numerze l	element ciągu Q_{RES} o numerze l	Uwaga 4.1.3.3., R4.1.3.
f_{dis} $f_{dis}(z, m_l)$	funkcja zakłóceń (funkcja częściowa)	$f_{dis}: Z \times M \rightarrow O$ funkcja częściowa $f_{dis}: Z \times M \rightarrow O$, przy czym jeżeli $f_{dis}(z_l, m_l) = o^l$, to: $z_l \in Z$ – zdarzenie elementarne polegające na wystąpieniu zakłócenia o pewnej wartości wpływu na wynik interakcji, $m_l \in M$ – czas interakcji o numerze l, a $o^l \in O$ – jakość dostarczonej usługi w ramach interakcji w ocenie usługobiorcy (rzeczywisty wynik interakcji).	Definicja 4.1.3.5., R4.1.3.
O	zbiór wyników interakcji	$O = Q$	Uwaga 4.1.3.4., R4.1.3.

O_{RES}	ciąg rzeczywistych wyników interakcji	$O_{RES}: L_I \to O$ gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru O i odpowiadają kolejnym rzeczywistym wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji	Definicja 4.1.3.6., R4.1.3.
o^l	rzeczywisty wynik interakcji o numerze l	element ciągu O_{RES} o numerze l	Uwaga 4.1.3.6., R4.1.3.
$\mathbb{A}$	rodzina wszystkich podzbiorów zbioru A	$\mathbb{A} = P(A)$	Definicja 4.1.4., R4.1.4.
f_{sel} $f_{sel}(A_x, m_l)$	funkcja wyboru usługodawcy (funkcja częściowa)	$f_{sel}: \mathbb{A}\ x\ M \to A$ gdzie $A_x \in \mathbb{A}$ – zbiór potencjalnych usługodawców, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – wybrany usługodawca (agent dostarczający usługę)	Definicja 4.1.4., R4.1.4.
T	zbiór możliwych wartości zaufania		Definicja 4.2.1.1., R4.2.1.
t_n	wartość zaufania	$t_n \in T$	Uwaga 4.2.1.1., R4.2.1.
t_{max}	maksymalna wartość zaufania	$t_{max} \in T$	Uwaga 4.2.1.2., R4.2.1.
t_{min}	minimalna wartość zaufania	$t_{min} \in T$	Uwaga 4.2.1.2., R4.2.1.
f_{trust} $f_{trust}(a_i, a_j, c_k, m_l)$	zaufanie (funkcja częściowa)	$f_{trust}: A \times A \times C \times M \to T$ przy czym jeżeli $f_{trust}(a_i, a_j, c_k, m_l) = t_n$, to $a_i \in A$ - agent ufający, $a_j \in A$ - agent zaufany, $c_k \in C$ – kontekst zaufania, $m_l \in M$ – czas (interakcji lub oceny zaufania), $t_n \in T$ – wartość zaufania.	Definicja 4.2.1.2., R4.2.1.
$t_{a_i \to a_j}^{c_k; m_l}$	wartość zaufania agenta a_i do agenta a_j w kontekście c_k w chwili m_l		R4.2.1.
$\vec{t}_{a_i \to a_j}^{C; m_l}$	wektor wartości zaufania agenta a_i do agenta a_j we wszystkich kontekstach w chwili m_l		R4.2.1.
$T_{a_i \to A}^{C; m_l}$	macierz wartości zaufania agenta a_i do wszystkich agentów we wszystkich kontekstach w chwili m_l		R4.2.1.
$\vec{t}_{a_i \to A}^{c_k; m_l}$	wektor wartości zaufania agenta a_i do wszystkich agentów w kontekście c_k w chwili m_l		R4.2.1.
$T_{A \to A}^{c_k; m_l}$	macierz wartości zaufania wszystkich agentów do		R4.2.1.

	wszystkich agentów w kontekście c_k w chwili m_l.		
P	zbiór możliwych wartości reputacji		Definicja 4.2.1.3., R4.2.1.
p_n	wartość reputacji	$p_n \in P$	Uwaga 4.2.1.3., R4.2.1.
p_{max}	maksymalna wartość reputacji	$p_{max} \in$ P	Uwaga 4.2.1.4., R4.2.1.
p_{min}	minimalna wartość reputacji	$p_{min} \in$ P	Uwaga 4.2.1.4., R4.2.1.
f_{rep} $f_{rep}(a_j, c_k, m_l)$	reputacja (funkcja częściowa)	$f_{rep}: A \times C \times M \to P$ przy czym $a_j \in A$ - agent obdarzony reputacją (zaufany), $c_k \in C$ – kontekst reputacji, $m_l \in M$ – czas (interakcji lub oceny reputacji), $p_n \in P$ – wartość reputacji	Definicja 4.2.1.4., R4.2.1.
$p_{a_j}^{c_k;m_l}$	wartość reputacji agenta a_j w kontekście c_k w chwili m_l		R4.2.1.
$\vec{p}_{a_j}^{C;m_l}$	wektor wartości reputacji agenta a_j we wszystkich kontekstach w chwili m_l		R4.2.1.
$\vec{p}_A^{c_k;m_l}$	wektor wartości reputacji wszystkich agentów w kontekście c_k w chwili m_l		R4.2.1.
$P_A^{C;m_l}$	macierz wartości reputacji wszystkich agentów we wszystkich kontekstach w chwili m_l		R4.2.1.
C_α	zbiór wszystkich możliwych kontekstów		Uwaga 4.2.2.1., R4.2.2.
C	zbiór kontekstów w systemie	$C \subset C_\alpha$ zawiera konteksty, dla których mogą być określone wartości zaufania lub reputacji lub dla których mogą być wydawane rekomendacje w danym systemie TRM	Definicja 4.2.2.2., R4.2.2.
c_k	kontekst k	$c_k \in C_\alpha$	Definicja 4.2.2.1., R4.2.2.
$r(U_k)$	kontekst dotyczący rekomendacji na temat świadczenia usług ze zbioru U_k		R4.2.2.
$r(c_k)$	kontekst dotyczący rekomendacji na temat kontekstu c_k		R4.2.2.

R	zbiór możliwych wartości rekomendacji		Definicja 4.2.3.1., R4.2.3.
r_n	wartość rekomendacji	$r_n \in R$	Uwaga 4.2.3.1., R4.2.3.
f_{reca_in} $f_{reca\_in}(a_j, a_p, c_k, m_l)$	funkcja rekomendacji wewnętrznej agenta (funkcja częściowa)	$f_{reca\_in}: A \times A \times C \times M \rightarrow R$ przy czym jeżeli $f_{reca\_in}(a_j, a_p, c_k, m_l) = r_n$, to $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $a_p \in A$ – agent, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji wewnętrznej agenta.	Definicja 4.2.3.2., R4.2.3.1.
f_{reca} $f_{reca}(a_i, a_j, a_p, c_k\ m_l)$	funkcja rekomendacji agenta (funkcja częściowa)	$f_{reca}: A \times A \times A \times C \times M \rightarrow R$ przy czym jeżeli $f_{reca}(a_i, a_j, a_p, c_k, m_l) = r_n$, to $a_i \in A$ – agent żądający rekomendacji (odbiorca rekomendacji), $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $a_p \in A$ – agent, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	Definicja 4.2.3.4., R4.2.3.1.
R^{AS}	zbiór rekomendacji agenta	zbiór krotek 6-elementowych $(a_i, a_j, a_p, c_k, m_l, r_n)$	Definicja 4.2.3.5., R4.2.3.1.
f_{reci_in} $f_{reci\_in}(a_j, a_p, c_k, m_l)$	funkcja rekomendacji wewnętrznej interakcji (funkcja częściowa)	$f_{reci\_in}: A \times I \times C \times M \rightarrow R$ przy czym jeżeli $f_{reci\_in}(a_j, i_p, c_k, m_l) = r_n$, to $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $i_p \in I$ – element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	Definicja 4.2.3.6., R4.2.3.2.
f_{reci} $f_{reci}(a_i, a_j, a_p, c_k, m_l)$	funkcja rekomendacji interakcji (funkcja częściowa)	$f_{reci}: A \times A \times I \times C \times M \rightarrow R$ przy czym jeżeli $f_{reci}(a_i, a_j, i_p, c_k, m_l) = r_n$, to $a_i \in A$ – agent żądający rekomendacji (odbiorca rekomendacji), $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $i_p \in I$ – element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	Definicja 4.2.3.8., R4.2.3.2.
R^{IS}	zbiór rekomendacji interakcji	zbiór krotek 6-elementowych $(a_i, a_j, i_p, c_k, m_l, r_n)$	Definicja 4.2.3.9., R4.2.3.2.
f_{sel_rec} $f_{sel\_rec}(c_l, a_k, m_l)$	funkcja wyboru dostawców rekomendacji (funkcja częściowa)	$f_{sel\_rec}: C \times A \times M \rightarrow \mathbb{A}$ *przy czym jeżeli* $f_{sel\_rec}(c_i, a_j, m_l) = A_p$, *to* $c_i \in C$ *– kontekst rekomendacji*, $a_k \in A$ *– podmiot rekomendacji,*, $m_l \in M$ *– czas*, $A_p \in \mathbb{A}$ *– zbiór dostawców rekomendacji.*	Definicja 4.2.3.10., R4.2.3.5.

$e_{R:l}$	żądanie rekomendacji o numerze l	uporządkowana 4-elementową krotka (a_i, a_p, c_k, m_l) lub (a_i, i_p, c_k, m_l), gdzie: $a_i \in A$ – żądający rekomendacji, $a_p \in A$ – przedmiot rekomendacji (agent) lub $i_p \in I$ – przedmiot rekomendacji (interakcja), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas pojawienia się żądania o numerze l	Definicja 4.2.3.11., R4.2.3.5.
$R^{S:m_z}$	zbiór rekomendacji wydanych do chwili m_z	$R^{S:m_z} \subseteq R^S$	Definicja 4.2.3.12., R4.2.3.8.
$r_{a_i \to a_j:a_p}^{c_k;m_l}$	wartość rekomendacji wydanej przez agenta a_i przesłanej do agenta a_j na temat agenta a_p w kontekście c_k w chwili m_l		R4.2.3.8.
$r_{a_i \to a_j:i_p}^{c_k;m_l}$	wartość rekomendacji wydanej przez agenta a_i przesłanej do agenta a_j na temat interakcji i_p w kontekście c_k w chwili m_l		R4.2.3.8.
$r_{a_i \to A:a_p}^{c_k;m_l}$	wartość rekomendacji publicznej (dostępnej dla wszystkich agentów), wydanej przez agenta a_i na temat agenta a_p w kontekście c_k w chwili m_l		R4.2.3.8.
$r_{a_i \to A:i_p}^{c_k;m_l}$	wartość rekomendacji publicznej (dostępnej dla wszystkich agentów), wydanej przez agenta a_i na temat interakcji i_p w kontekście c_k w chwili m_l		R4.2.3.8.
f_{obs} $f_{obs}(a_i, a_j, a_k, u_l, m_n)$	obserwacja (funkcja częściowa)	$f_{obs}: A \times A \times A \times U \times M \to O$ przy czym jeżeli $f_{obs}(a_i, a_j, a_k, u_l, m_n) = o^n$, to $a_i \in A$ – agent obserwujący, $a_j \in A$ – agent żądający usługi, $a_k \in A$ – agent dostarczający usługę, $u_l \in U$ – usługa, $m_n \in M$ – czas, $o^n \in O$ – rzeczywisty wynik tej interakcji pomiędzy agentami a_j i a_k zaobserwowany przez agenta a_i.	Definicja 4.2.4., R4.2.4.
E_q	efektywność środowiska bez zakłóceń	$E_q = \frac{\sum_{i=1}^{l} q^i}{l * q_{max}}$	Definicja 5.1.1.1., R5.1.1.
E	efektywność środowiska	$E = \frac{\sum_{i=1}^{l} o^i}{l * q_{max}}$	Definicja 5.1.1.2., R5.1.1.

$E^{(n)}$	efektywność chwilowa n	$E^{(n)} = \frac{\sum_{i=l-n+1}^{l} o^i}{n * q_{max}}$	Definicja 5.1.1.3., R5.1.1.
E^{ideal}	idealna efektywność	$E^{ideal} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{max}^{u_{int:i}}}$	Definicja 5.1.1.4., R5.1.1.
$q_{max}^{u_l}$	maksymalna jakość usługi u_l z jaką może być świadczona	$q_{max}^{u_l} = q_{a_k}^{u_l} : \forall a_j \in A_{P:u_l} : q_{a_j}^{u_l} \leq q_{a_k}^{u_l}$	R5.1.1.
$u_{int:l}$	usługa, która była świadczona w ramach l – tej interakcji		R5.1.1.
$q_{max}^{u_{int:i}}$	maksymalna jakość usługi (w całym środowisku), która była świadczona podczas i – tej interakcji		Definicja 5.1.1.4., R5.1.1.
E^{adv}	zaawansowana efektywność	$E^{adv} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{a_{int:i}}^{u_{int:i}}}$ $E^{adv} \geq E^{ideal} \geq E$	Definicja 5.1.1.5., R5.1.1.
G	zysk efektywności	$G = E_{+TRM} - E_0$ E_{+TRM} - efektywność środowiska, w którym działa określony system zarządzania zaufaniem E_0 - efektywność środowiska bez systemu zarządzania zaufaniem, przy założeniu, że w obydwu przypadkach atakujący zachowują się dokładnie w ten sam sposób – stosują działania pozwalające maksymalnie obniżyć efektywność środowiska w momencie korzystania z systemu zarządzania zaufaniem	Definicja 5.1.2.2., R5.1.2.
G_A	zysk absolutny efektywności	$G_A = E_{+TRM} - E_0$, E_{+TRM} - efektywność środowiska, w którym działa określony system zarządzania zaufaniem E_0 - efektywność środowiska bez systemu zarządzania zaufaniem, przy założeniu, że w obydwu przypadkach atakujący – stosują działania pozwalające maksymalnie obniżyć efektywność środowiska, tzn. tak dobierają swoje działania aby maksymalnie obniżyć efektywność zarówno w przypadku kiedy system zarządzania zaufaniem jest używany, jak i wtedy gdy nie jest	Definicja 5.1.2.3., R5.1.2.
$t_{A\to A}^{c_k;m_l}$	globalne średnie zaufanie w kontekście $\boldsymbol{c_k}$	$\overline{t_{A\to A}^{c_k;m_l}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n}\sum_{j=1}^{n} 1}$ przy czym i, j są takie że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.3.1., R5.1.3.
$t_{A_B\to A}^{c_k;m_l}$	globalne średnie zaufanie agentów rzetelnych do wszystkich agentów w kontekście c_k	$\overline{t_{A_B\to A}^{c_k;m_l}} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n} 1}$ przy czym i, j jest takie, że: $a_i \in A_B, a_j \in A, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone	Definicja 5.1.3.2., R5.2.3.
$t_{A_B\to A_B}^{c_k;m_l}$	globalne średnie zaufanie agentów rzetelnych do	$\overline{t_{A_B\to A_B}^{c_k;m_l}} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n_B} 1}$	Definicja 5.1.3.3., R5.1.3.

	agentów rzetelnych w kontekście c_k	przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_B, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone	
$t_{A_B \to A_M}^{c_k;m_l}$	globalne średnie zaufanie agentów rzetelnych do agentów złośliwych w kontekście c_k	$\overline{t_{A_B \to A_M}^{c_k;m_l}} = \frac{\sum_{i=1}^{n_B} \sum_{j=1}^{n_M} t_{a_i \to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B} \sum_{j=1}^{n_M} 1}$ przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_M$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone	Definicja 5.1.3.4., R5.1.3.
$p_A^{c_k;m_l}$	globalna średnia reputacja agentów w kontekście c_k	$\overline{p_A^{c_k;m_l}} = \frac{\sum_{j=1}^{n} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n} 1}$ przy czym j jest takie, że $a_j \in A$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$)	Definicja 5.1.3.5., R5.1.3.
$p_{A_B}^{c_k;m_l}$	globalna średnia reputacja agentów rzetelnych w kontekście c_k	$\overline{p_{A_B}^{c_k;m_l}} = \frac{\sum_{j=1}^{n_B} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_B} 1}$ przy czym j jest takie, że $a_j \in A_B$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$)	Definicja 5.1.3.6., R5.1.3.
$p_{A_M}^{c_k;m_l}$	globalna średnia reputacja agentów złośliwych w kontekście c_k	$\overline{p_{A_M}^{c_k;m_l}} = \frac{\sum_{j=1}^{n_M} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_M} 1}$ przy czym j jest takie, że $a_j \in A_M$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$)	Definicja 5.1.3.7., R5.1.3.
$t_{A \to a_j}^{c_k;m_l}$	średnie zaufanie do agenta a_j w kontekście c_k	$t_{A \to a_j}^{c_k;m_l} = \frac{\sum_{i=1}^{n} t_{a_i \to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n} 1}$ przy czym i jest takie, że: $a_i \in A,\ i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone	Definicja 5.1.4., R5.1.4.
$I^{P:a_j}$	zbiór interakcji, w których agent a_j jest usługodawcą	$I^{P:a_j} \subseteq I$	R5.1.5.
$I^{R:a_i}$	zbiór interakcji, w których agent a_i jest usługobiorcą	$I^{R:a_i} \subseteq I$	R5.1.5.
$I^{R:a_i,P:a_j}$	zbiór interakcji, w których agent a_i jest usługobiorcą, a agent a_j jest usługodawcą	$I^{R:a_i,P:a_i} \subseteq I$, $I^{R:a_i,P:a_i} \subseteq I^{P:a_j}$, $I^{R:a_i,P:a_i} \subseteq I^{R:a_i}$	R5.1.5.
$i_k^{P:a_j}$	k-ty element zbioru interakcji, w których agent a_j jest usługodawcą		R5.1.5.
$i_k^{R:a_i}$	k-ty element zbioru interakcji, w których agent a_i jest usługobiorcą		R5.1.5.
$i_k^{R:a_i,P:a_j}$	k-ty element zbioru interakcji, w których agent a_j jest usługodawcą a agent a_i jest usługobiorcą		R5.1.5.
$l_I^{P:a_j}$	liczba interakcji, w których agent a_j jest usługodawcą	$l_I^{P:a_j} = \lvert I^{P:a_j} \rvert$	R5.1.5.

$l_I^{R:a_i}$	liczba interakcji, w których agent a_i jest usługobiorcą	$l_I^{R:a_i} = \lvert I^{P:a_i} \rvert$	R5.1.5.
$l_I^{R:a_i,P:a_j}$	liczba interakcji, w których agent a_j jest usługodawcą, a agent a_i usługobiorcą	$l_I^{R:a_i,P:a_j} = \left\lvert I^{R:a_i,P:a_j} \right\rvert$	R5.1.5.
$o_\Sigma^{P:a_j}$	suma rzeczywistych wyników interakcji, w których agent a_j był usługodawcą		R5.1.5.
$l_I^{A_B,A_B}$	liczba interakcji agentów rzetelnych z agentami rzetelnymi	$l_I^{A_B,A_B} = \sum_{i,j:a_i,a_j \in A_B, i \neq j} l_I^{R:a_i,P:a_j}$	Definicja 5.1.5.1., R5.1.5.
$l_I^{A_B,A_M}$	liczba interakcji agentów rzetelnych z agentami złośliwymi	$l_I^{A_B,A_M} = \sum_{i:a_i \in A_B, j:a_j \in A_M} l_I^{R:a_i,P:a_j} + \sum_{i:a_i \in A_M, j:a_j \in A_B} l_I^{R:a_i,P:a_j}$	Definicja 5.1.5.2., R5.1.5.
$o_E^{P:a_j}$	średni rzeczywisty wynik usług świadczonych przez agenta a_j	$\forall_{a_j \in A_P, l_I^{P:a_j} > 0} \; o_E^{P:a_j} = \frac{o_\Sigma^{P:a_j}}{l_I^{P:a_j}}$	Definicja 5.1.5.3., R5.1.5.
$\overline{l_I^{P:A}}$	średnia liczba interakcji, w których agenci byli usługodawcami	$\overline{l_I^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} l_I^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$	Definicja 5.1.5.4., R5.1.5.
$\overline{o_E^{P:A}}$	średnia rzeczywista jakość usług	$\overline{o_E^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} o_E^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$	Definicja 5.1.5.5., R5.1.5

ZAŁĄCZNIK 3 – WYBRANE POJĘCIA STOSOWANE W PRACY

Niniejszy załącznik zawiera opis wybranych pojęć, które zostały zastosowane w pracy.

Pojęcie	**Objaśnienie**	**Miejsce wprowadzenia i wyjaśnienia w rozprawie**
system wieloagentowy	system złożony z komunikujących i współpracujących między sobą agentów, realizujących określone cele	Definicja 1, R2.1
środowisko	system wieloagentowy, składający się z agentów wchodzących w interakcje polegające na świadczeniu określonych usług	Definicja 2., R2.1
środowisko	jest uporządkowaną 5-elementową krotką (5-ką) (A, U, E, F, M), gdzie A jest zbiorem agentów (łącznie z ich charakterystyką), U jest zbiorem usług, E jest zbiorem żądań, F jest zbiorem funkcji częściowych określonych w środowisku $F = \{f_{sel}, f_{int}, f_{dis}\}$, a M jest zbiorem czasów interakcji	Definicja 4.1.5., R4.1.5
mechanizmy twardego bezpieczeństwa	są związane z indywidualnymi metodami ochrony, których skuteczność opiera się na ściśle określonych regułach działania i ugruntowanych podstawach matematycznych	Definicja 3., R2.3
mechanizmy miękkiego bezpieczeństwa	są związane z kolektywnymi metodami ochrony, które zakładają, że w systemie może znaleźć się intruz lub nieoczekiwanie działający komponent. Zadaniem mechanizmów miękkiego bezpieczeństwa jest wykrycie tego rodzaju zdarzeń lub adwersarzy i uniemożliwienie im spowodowania szkód dla całego systemu	Definicja 4., R2.3
zaufanie i relacja zaufania	**Zaufanie** charakteryzuje relację pomiędzy parą agentów, dotyczącą określonego kontekstu, będącą oceną agenta ufającego co do rzetelności zachowania agenta zaufanego. Wobec tego **relacja zaufania** jest uporządkowaną 5-elementową krotką (5-ką) (A_1, A_2, v, c, m), gdzie A_1 jest agentem ufającym, A_2 agentem zaufanym, v jest pewną wartością zaufania z określonego zbioru możliwych wartości, charakteryzującą moc relacji zaufania, c charakteryzuje kontekst relacji zaufania, a m jest czasem, w którym dana relacja zachodzi	Definicja 5., R2.4
zbiór możliwych wartości zaufania	T jest zbiorem możliwych wartości zaufania określonych przez system TRM	Definicja 4.2.1.1., R4.2.1,
zaufanie (funkcja częściowa)	funkcja częściowa $f_{trust}: A \times A \times C \times M \to T$, przy czym jeżeli $f_{trust}(a_i, a_j, c_k, m_l) = t_n$, to $a_i \in A$ - agent ufający, $a_j \in A$ - agent zaufany, $c_k \in C$ – kontekst zaufania, $m_l \in M$ – czas (interakcji lub oceny zaufania), $t_n \in T$ – wartość zaufania	Definicja 4.2.1.2., R4.2.1,
zaufanie (funkcja częściowa) – uogólnienie	Niech $\mathbb{A}$ będzie zbiorem potęgowym zbioru A bez zbioru pustego ($\mathbb{A} = P(A) - \{\emptyset\}$). Wtedy każdy z elementów rodziny $\mathbb{A}$ jest pewnym podzbiorem zbioru A i można go utożsamiać z pewną grupą agentów. Zaufanie może być określone jako funkcja częściowa $f_{trust}: \mathbb{A} \times \mathbb{A} \times C \times M \to T$, przy czym jeżeli $f_{trust}(A_i, A_j, c_k, m_l) = t_n$, to $A_i \in \mathbb{A}$ – zbiór agentów ufających, $A_j \in \mathbb{A}$ – zbiór agentów zaufanych, $c_k \in C$ – kontekst zaufania, $m_l \in M$ – czas (interakcji lub oceny zaufania), $t_n \in T$ – wartość zaufania.	Definicja 4.2.1.2'., R4.2.1,
reputacja i opinia o reputacji	**Reputacja** charakteryzuje opinię o danym agencie, dotyczącą określonego kontekstu, będącą miarą oceny grupy agentów co do rzetelności zachowania tego agenta. Wobec tego **opinia o reputacji** jest uporządkowaną 4-elementową krotką (4-ką) (A_2, v, c, m), gdzie A_2 jest agentem zaufanym, v jest pewną wartością reputacji z określonego zbioru możliwych wartości, charakteryzującą poziom reputacji agenta,	Definicja 6., R2.4

	c charakteryzuje kontekst reputacji, a m jest czasem, w którym dana relacja zachodzi.	
zbiór możliwych wartości reputacji	P jest zbiorem możliwych wartości reputacji określonych przez system TRM.	Definicja 4.2.1.3., R4.2.1,
reputacja (funkcja częściowa)	funkcja częściowa $f_{rep}: A \times C \times M \to P$, przy czym jeżeli $f_{rep}(a_j, c_k, m_l) = p_n$, to $a_j \in A$ - agent obdarzony reputacją (zaufany), $c_k \in C$ – kontekst reputacji, $m_l \in M$ – czas (interakcji lub oceny reputacji), $p_n \in P$ – wartość reputacji.	Definicja 4.2.1.4., R4.2.1,
reputacja (funkcja częściowa) – uogólnienie	Niech $\mathbb{A}$ będzie zbiorem potęgowym zbioru A bez zbioru pustego ($\mathbb{A} = P(A) - \{\emptyset\}$). Wtedy każdy z elementów rodziny $\mathbb{A}$ jest pewnym podzbiorem zbioru A i można go utożsamiać z pewną grupą agentów. Reputacja może być określona jako funkcja częściowa $f_{rep}: \mathbb{A} \times C \times M \to P$, przy czym jeżeli $f_{rep}(A_j, c_k, m_l) = p_n$, to $A_j \in \mathbb{A}$ – zbiór agentów obdarzonych reputacją (zaufanych), $c_k \in C$ – kontekst reputacji, $m_l \in M$ – czas (interakcji lub oceny reputacji), $p_n \in P$ – wartość reputacji.	Definicja 4.2.1.4'., R4.2.1,
system TRM	system, działający w ramach określonego środowiska, który na bazie informacji o ocenie rzetelności wymienianej pomiędzy agentami lub na bazie informacji o ocenie rzetelności dostarczanych do i przez zaufaną trzecią stronę, wspomaga decyzję agenta co do wyboru przyszłego partnera interakcji lub jakości świadczonej usługi, w celu maksymalizacji własnej użyteczności agenta lub środowiska.	Definicja 7., R2.5
atak na system TRM	wszelkie działania podejmowane przez agenta lub grupę agentów, które mają na celu zaburzenie lub zmianę rezultatu obliczeń miar zaufania lub reputacji, lub zmianę rezultatu procesu podejmowania decyzji przez innych agentów w oparciu o obliczone miary zaufania lub reputacji.	Definicja 8., R2.8
wiarygodność	cechą systemów TRM, będącą miarą poziomu odporności na ataki. System TRM jest wiarygodny, jeżeli nie istnieje taki znany atak na system TRM, który jest w stanie istotnie zaburzyć (zmienić) decyzje podejmowane przez agenty w oparciu o ten system TRM.	Definicja 9. i 10., R2.9
żądanie	jest uporządkowaną 3-elementową krotką $(a_i\ u_k, m_l)$, gdzie: $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana usługa, $m_l \in M$ – czas pojawienia się żądania o numerze l.	Definicja 4.1.2., R4.1.2.
funkcja interakcji	funkcja częściowa $f_{int}: A \times U \times M \times A \to Q$, przy czym jeżeli $f_{int}(a_i, u_k, m_l, a_j) = q^l$, to $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana (świadczona) usługa, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę), a $q^l \in Q$ – jakość dostarczonej usługi w ramach l-tej interakcji (wynik tej interakcji).	Definicja 4.1.3.1., R4.1.3
zbiór interakcji	dziedzina funkcji interakcji	Definicja 4.1.3.2., R4.1.3
interakcja	element zbioru interakcji o numerze l (argument funkcji interakcji o numerze l), czyli krotka: (a_i, u_k, m_l, a_j), przy czym $a_i \in A$ – usługobiorca (agent żądający usługi), $u_k \in U$ – żądana (świadczona) usługa, $m_l \in M$ – czas rozpoczęcia interakcji o numerze l, $a_j \in A$ – usługodawca (agent dostarczający usługę).	Definicja 4.1.3.3., R4.1.3
ciąg wyników interakcji	to ciąg $Q_{RES}: L_I \to Q$, gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru Q i odpowiadają kolejnym wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji.	Definicja 4.1.3.4., R4.1.3
funkcja zakłóceń	funkcja częściowa $f_{dis}: Z \times M \to O$, przy czym jeżeli $f_{dis}(z_l, m_l) = o^l$, to: $z_l \in Z$ – zdarzenie elementarne polegające na wystąpieniu zakłócenia o pewnej wartości wpływu na wynik interakcji, $m_l \in M$ – czas interakcji o numerze l, a $o^l \in O$ – jakość dostarczonej usługi w ramach interakcji w ocenie usługobiorcy (rzeczywisty wynik interakcji).	Definicja 4.1.3.5., R4.1.3
ciąg rzeczywistych	ciąg $O_{RES}: L_I \to O$, gdzie indeksy są numerami kolejnych interakcji, a wartości ciągu są ze zbioru O i odpowiadają kolejnym rzeczywistym	Definicja 4.1.3.6., R4.1.3

wyników interakcji	wynikom interakcji, uszeregowanym według czasu rozpoczęcia interakcji	
funkcja wyboru usługodawcy	Niech $\mathbb{A}$ będzie rodziną skończoną wszystkich podzbiorów zbioru A ($\mathbb{A} = P(A)$, $\mathbb{A}$ jest zbiorem potęgowym zbioru A). Wtedy każdy z elementów rodziny $\mathbb{A}$ jest pewnym podzbiorem zbioru A i można go utożsamiać ze zbiorem potencjalnych usługodawców. Funkcją wyboru usługodawcy jest funkcja częściowa $f_{sel}: \mathbb{A}\ x\ M \rightarrow A$, która dla dowolnego $X \in \mathbb{A}$ spełnia warunek $f_{sel}(X, m) \in X$.	Definicja 4.1.4., R4.1.4
funkcja wyboru usługodawcy – uogólnienie	. Niech $\overline{\overline{\mathbb{A}}}$ będzie zbiorem potęgowym zbioru $\mathbb{A}$ ($\overline{\overline{\mathbb{A}}} = P(\mathbb{A}) = P(P(A))$, czyli $\mathbb{A}$ jest zbiorem potęgowym zbioru A, a $\overline{\overline{\mathbb{A}}}$ jest zbiorem potęgowym zbioru $\mathbb{A}$). Wtedy każdy z elementów $\overline{\overline{\mathbb{A}}}$ jest pewnym podzbiorem zbioru $\mathbb{A}$. Istnieje podzbiór $\overline{\overline{\mathbb{A}'}} \subset \overline{\overline{\mathbb{A}}}$, który zawera te elementy zbioru $\overline{\overline{\mathbb{A}}}$, (czyli te zbiory agentów) które mogą kolektywnie świadczyć usługę. Zbiór $\overline{\overline{\mathbb{A}'}}$ jest więc zbiorem potencjalnych usługodawców. Funkcją wyboru usługodawcy jest funkcja częściowa $f_{sel}: \overline{\overline{\mathbb{A}'}}\ x\ M \rightarrow \mathbb{A}$, która dla dowolnego $X \in \overline{\overline{\mathbb{A}'}}$ spełnia warunek $f_{sel}(X, m) \in X$.	Definicja 4.1.4'., R4.1.4
kontekst	Kontekst $c_k \in C_\alpha$ określa to czego dotyczy rekomendacja lub miara zaufania lub reputacji. Kontekstem jest: • świadczenie usług z dowolnego niepustego podzbioru zbioru wszystkich usług: $c_k = U_k$, gdzie $U_k \subseteq U$, $U_k \neq \emptyset$, $U = \{u_1, \dots, u_l\}$; lub • rekomendacja dotycząca dowolnego kontekstu: $c_l = r(c_k)$; lub • dowolny zbiór kontekstów: $c_m = \{c_k, c_l, \dots\}$.	Definicja 4.2.2.1., R4.2.2
zbiór kontekstów	Zbiór kontekstów C zawiera konteksty, dla których mogą być określone wartości zaufania lub reputacji lub dla których mogą być wydawane rekomendacje w danym systemie TRM. Zbiór kontekstów C jest skończonym podzbiorem zbioru wszystkich możliwych kontekstów: $C \subset C_\alpha$.	Definicja 4.2.2.2., R4.2.2
zbiór możliwych wartości rekomendacji	R jest zbiorem możliwych wartości rekomendacji określonych przez system TRM.	Definicja 4.2.3.1., R4.2.2
funkcja rekomendacji wewnętrznej agenta	funkcja częściowa $f_{reca\_in}: A \times A \times C \times M \rightarrow R$, przy czym jeżeli $f_{reca\_in}(a_j, a_p, c_k, m_l) = r_n$, to $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $a_p \in A$ – agent, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji wewnętrznej agenta.	Definicja 4.2.3.2., R4.2.3.
rekomendacja wewnętrzna agenta	krotka 5-elementowa $(a_j, a_p, c_k, m_l, r_n)$, utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji wewnętrznej agenta o jej wartość.	Definicja 4.2.3.3., R4.2.3.
funkcja rekomendacji agenta	funkcja częściowa $f_{reca}: A \times A \times A \times C \times M \rightarrow R$, przy czym jeżeli $f_{reca}(a_i, a_j, a_p, c_k, m_l) = r_n$, to $a_i \in A$ – agent żądający rekomendacji (odbiorca rekomendacji), $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $a_p \in A$ – agent, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	Definicja 4.2.3.4., R4.2.3.1,
rekomendacja agenta	krotka 6-elementowa $(a_i, a_j, a_p, c_k, m_l, r_n)$, utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji agenta o jej wartość. Zbiór takich krotek 6-elementowych jest zbiorem rekomendacji agenta R^{AS}.	Definicja 4.2.3.5., R4.2.3.1,
funkcja rekomendacji	funkcja częściowa $f_{reci\_in}: A \times I \times C \times M \rightarrow R$, przy czym jeżeli $f_{reci\_in}(a_j, i_p, c_k, m_l) = r_n$, to $a_j \in A$ – agent dostarczający	Definicja 4.2.3.6., R4.2.3.2,

wewnętrznej interakcji	rekomendację (wydawca rekomendacji), $i_p \in I$ – element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	
rekomendacja wewnętrzna interakcji	krotka 5-elementowa $(a_j, i_p, c_k, m_l, r_n)$, utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji wewnętrznej interakcji o jej wartość.	Definicja 4.2.3.7., R4.2.3.2,
funkcja rekomendacji interakcji	funkcja częściowa $f_{reci}: A \times A \times I \times C \times M \rightarrow R$, przy czym jeżeli $f_{reci}(a_i, a_j, i_p, c_k, m_l) = r_n$, to $a_i \in A$ – agent żądający rekomendacji (odbiorca rekomendacji), $a_j \in A$ – agent dostarczający rekomendację (wydawca rekomendacji), $i_p \in I$ – element ze zbioru interakcji, którego dotyczy rekomendacja (przedmiot rekomendacji), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas, $r_n \in R$ – wartość rekomendacji.	Definicja 4.2.3.8., R4.2.3.2
rekomendacja interakcji	krotka 6-elementowa $(a_i, a_j, i_p, c_k, m_l, r_n)$, utworzona przez poszerzenie krotki będącej argumentami funkcji rekomendacji interakcji o jej wartość. Zbiór takich krotek 6-elementowych jest zbiorem rekomendacji interakcji R^{IS}.	Definicja 4.2.3.9., R4.2.3.2
rekomendacja	Rekomendacja to rekomendacja agenta lub rekomendacja interakcji	Termin 4.2.3.1., R4.2.3.3.
zbiór rekomednacji	Zbiorem rekomendacji R^S jest suma zbioru rekomendacji agenta i zbioru rekomendacji interakcji, $R^S = R^{AS} \cap R^{IS}$.	Termin 4.2.3.2., R4.2.3.3.
funkcja wyboru dostawców rekomendacji	Niech $\mathbb{A}$ będzie zbiorem potęgowym zbioru A ($\mathbb{A} = P(A)$). Wtedy każdy z elementów $\mathbb{A}$ jest pewnym podzbiorem zbioru A i można go utożsamiać ze zbiorem potencjalnych dostawców rekomendacji. Funkcją wyboru dostawców rekomendacji jest funkcja częściowa $f_{sel\_rec}: C \times A \times M \rightarrow \mathbb{A}$, przy czym jeżeli $f_{sel\_rec}(c_i, a_j, m_l) = A_p$, to $c_i \in C$ – kontekst rekomendacji, $a_k \in A$ – podmiot rekomendacji,, $m_l \in M$ – czas, $A_p \in \mathbb{A}$ – zbiór dostawców rekomendacji.	Definicja 4.2.3.10., R4.2.3.5
żądanie rekomendacji	Żądanie rekomendacji $e_{R:l}$ jest uporządkowaną 4-elementową krotką $(a_i, a_p\ c_k, m_l)$ lub $(a_i, i_p\ c_k, m_l)$, gdzie: $a_i \in A$ – żądający rekomendacji, $a_p \in A$ – przedmiot rekomendacji (agent) lub $i_p \in I$ – przedmiot rekomendacji (interakcja), $c_k \in C$ – kontekst rekomendacji, $m_l \in M$ – czas pojawienia się żądania o numerze l.	Definicja 4.2.3.11., R4.2.3.5
zbiór rekomendacji wydanych w systemie TRM do chwili m_z	Zbiorem rekomendacji wydanych w systemie TRM do chwili m_z jest zbiór $R^{S:m_z}$, który zawiera wszystkie rekomendacje wydane w systemie TRM do chwili m_z (dla których $m_l \leq m_z, m_l \in M$ – czas); zbiór $R^{S:m_z}$ jest podzbiorem zbioru rekomendacji R^S, tj. $R^{S:m_z} \subseteq R^S$.	Definicja 4.2.3.12., R4.2.3.8
algorytm oceny rekomendacji	Algorytm oceny rekomendacji jest sposobem postępowania w celu aktualizacji wartości zaufania do agenta, który wydał rekomendację (lub reputacji tego agenta).	Definicja 4.2.3.13., R4.2.3.10
obserwacja	funkcja częściowa $f_{obs}: A \times A \times A \times U \times M \rightarrow O$, przy czym jeżeli $f_{obs}(a_i, a_j, a_k, u_l, m_n) = o^n$, to $a_i \in A$ – agent obserwujący, $a_j \in A$ – agent żądający usługi, $a_k \in A$ – agent dostarczający usługę, $u_l \in U$ – usługa, $m_n \in M$ – czas, $o^n \in O$ – rzeczywisty wynik tej interakcji pomiędzy agentami a_j i a_k zaobserwowany przez agenta a_i	Definicja 4.2.4., R4.2.4.
algorytm oceny interakcji i agenta	Algorytm oceny interakcji i agenta jest sposobem postępowania w celu aktualizacji wartości zaufania do agenta, który uczestniczył w interakcji (lub reputacji tego agenta).	Definicja 4.2.6., R4.2.6.
środowisko z działającym systemem TRM	Środowisko z działającym systemem TRM jest uporządkowaną 9-elementową krotką (9-ką) $(A', U, E, F', M, T, P, C, R)$, gdzie A' jest zbiorem agentów (łącznie z ich charakterystyką), U jest zbiorem usług, E jest zbiorem żądań, F' jest zbiorem funkcji częściowych określonych w środowisku z systemem TRM, M jest zbiorem czasów interakcji, T	Definicja 4.2.8., R4.2.8.

	jest zbiorem wartości zaufania, P jest zbiorem wartości reputacji, C jest zbiorem kontekstów, a R jest zbiorem rekomendacji.	
zachowanie jednostkowe	zachowanie agenta związane z obsługą jednego żądania świadczenia usług	Definicja 4.3.1.1., R4.3.1.
atak elementarny	wyznaczenie wyniku przynajmniej jednej z funkcji częściowych w sposób niezgodny z tym określonym w środowisku lub systemie TRM lub niepoprawna identyfikacja agenta	Definicja 4.3.1.2., R4.3.1.
efektywność środowiska bez zakłóceń (E_q)	stosunek sumy wyników interakcji do iloczynu liczby wszystkich interakcji i maksymalnej jakości usług w środowisku: $$E_q = \frac{\sum_{i=1}^{l} q^i}{l * q_{max}}$$	Definicja 5.1.1.1., R5.1.1.
efektywność środowiska (E)	stosunek sumy rzeczywistych wyników interakcji do iloczynu liczby wszystkich interakcji i maksymalnej jakości usług w środowisku: $$E = \frac{\sum_{i=1}^{l} o^i}{l * q_{max}}$$	Definicja 5.1.1.2., R5.1.1.
efektywność chwilowa n ($E^{(n)}$)	efektywność środowiska biorąca pod uwagę jedynie n ostatnich interakcji: $$E^{(n)} = \frac{\sum_{i=l-n+1}^{l} o^i}{n * q_{max}}$$	Definicja 5.1.1.3., R5.1.1.
idealna efektywność (E^{ideal})	jest równa stosunkowi sumy rzeczywistych wyników interakcji do sumy maksymalnej jakości usług w całym środowisku świadczonych przez agenty w kolejnych interakcjach: $$E^{ideal} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{max}^{u_{int:i}}}$$ przy czym $q_{max}^{u_{int:i}}$ to maksymalna jakość usługi (w całym środowisku), która była świadczona podczas i – tej interakcji.	Definicja 5.1.1.4., R5.1.1.
zaawansowana efektywność (E^{adv})	jest równa stosunkowi sumy rzeczywistych wyników interakcji do sumy maksymalnej jakości usług danego agenta świadczonych w kolejnych interakcjach: $$E^{adv} = \frac{\sum_{i=1}^{l} o^i}{\sum_{i=1}^{l} q_{a_{int:i}}^{u_{int:i}}}$$ przy czym $q_{a_{int:i}}^{u_{int:i}}$ to maksymalna jakość usługi świadczonej przez agenta, który był usługodawcą podczas i – tej interakcji.	Definicja 5.1.1.5., R5.1.1.
najbardziej efektywny atak	taki sposób działania atakujących (podejmowania decyzji w odniesieniu do aspektów wymienionych w modelu ataku), który pozwala zminimalizować efektywność środowiska.	Definicja 5.1.2.1., R5.1.2.
zysk efektywności (G)	różnica pomiędzy efektywnością środowiska, w którym działa określony system zarządzania zaufaniem (E_{+TRM}), a efektywnością środowiska bez systemu zarządzania zaufaniem (E_0), przy założeniu, że w obydwu przypadkach atakujący zachowują się dokładnie w ten sam sposób – stosują działania pozwalające maksymalnie obniżyć efektywność środowiska w momencie korzystania z systemu zarządzania zaufaniem: $$G = E_{+TRM} - E_0$$	Definicja 5.1.2.2., R5.1.2.
zysk absolutny efektywności (G_A)	różnica pomiędzy efektywnością środowiska, w którym działa określony system zarządzania zaufaniem (E_{+TRM}) i efektywnością środowiska bez systemu zarządzania zaufaniem ($E_{0'}$) przy założeniu, że w obydwu przypadkach atakujący stosują najbardziej efektywny atak, tj. działania pozwalające maksymalnie obniżyć efektywność środowiska, tzn. tak dobierają swoje działania aby maksymalnie obniżyć efektywność zarówno w przypadku kiedy system zarządzania zaufaniem jest używany, jak i wtedy gdy nie jest (mogą stosować różne strategie działania w tych dwóch przypadkach): $$G_A = E_{+TRM} - E_{0'}$$	Definicja 5.1.2.3., R5.1.2.

globalne średnie zaufanie w kontekście c_k	suma wartości zaufania w kontekście c_k pomiędzy każdą uporządkowaną parą agentów, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania wszystkich agentów do wszystkich agentów ($A \to A$) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$t_{A\to A}^{c_k;m_l} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n}\sum_{j=1}^{n} 1}$$ przy czym i,j są takie że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.3.1., R5.1.3.
globalne średnie zaufanie agentów rzetelnych do wszystkich agentów w kontekście c_k	suma wartości zaufania w kontekście c_k pomiędzy agentem rzetelnym, a każdym innym agentem, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do wszystkich agentów ($A_B \to A$) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$t_{A_B\to A}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n} 1}$$ przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.3.2., R5.1.3.
globalne średnie zaufanie agentów rzetelnych do agentów rzetelnych w kontekście c_k	suma wartości zaufania w kontekście c_k pomiędzy każdą uporządkowaną parą agentów rzetelnych, dla których to zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do agentów rzetelnych ($A_B \to A_B$) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$t_{A_B\to A_B}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n_B} 1}$$ przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_B, i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.3.3., R5.1.3.
globalne średnie zaufanie agentów rzetelnych do agentów złośliwych w kontekście c_k	suma wartości zaufania w kontekście c_k każdego agenta rzetelnego do każdego agenta złośliwego, o ile pomiędzy tymi agentami zaufanie jest określone, podzielona przez iloczyn liczby takich par agentów i maksymalnej wartości zaufania. Globalna średnia wartość zaufania agentów rzetelnych do agentów złośliwych ($A_B \to A_M$) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$t_{A_B\to A_M}^{c_k;m_l} = \frac{\sum_{i=1}^{n_B}\sum_{j=1}^{n_M} t_{a_i\to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n_B}\sum_{j=1}^{n_M} 1}$$ przy czym i,j jest takie, że: $a_i \in A_B, a_j \in A_M$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.3.4., R5.1.3
globalna średnia reputacja agentów w kontekście c_k	suma wartości reputacji w kontekście c_k wszystkich agentów, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości reputacji. Globalna średnia reputacja wszystkich agentów (A) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$p_A^{c_k;m_l} = \frac{\sum_{j=1}^{n} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n} 1}$$ przy czym j jest takie, że $a_j \in A$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).	Definicja 5.1.3.5., R5.1.3
globalna średnia reputacja agentów	to suma wartości reputacji w kontekście c_k agentów rzetelnych, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości	Definicja 5.1.3.6., R5.1.3

rzetelnych w kontekście c_k	reputacji. Globalna średnia reputacja agentów rzetelnych (A_B) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$p_{A_B}^{c_k;m_l} = \frac{\sum_{j=1}^{n_B} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_B} 1}$$ przy czym j jest takie, że $a_j \in A_B$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).	
globalna średnia reputacja agentów złośliwych w kontekście c_k	suma wartości reputacji w kontekście c_k agentów złośliwych, o ile ta wartość jest określona przez funkcję częściową f_{rep}, podzielona przez iloczyn liczby takich wartości reputacji i maksymalnej wartości reputacji. Globalna średnia reputacja agentów złośliwych (A_M) w kontekście c_k w chwili m_l jest wyrażona wzorem: $$p_{A_M}^{c_k;m_l} = \frac{\sum_{j=1}^{n_M} p_{a_j}^{c_k;m_l}}{p_{max} * \sum_{j=1}^{n_M} 1}$$ przy czym j jest takie, że $a_j \in A_M$ oraz reputacja agenta a_j jest określona ($\exists f_{rep}(a_j, c_k, m_l)$).	Definicja 5.1.3.7., R5.1.3
średnie zaufanie do agenta a_j w kontekście c_k	suma wartości zaufania w kontekście c_k każdego agenta do agenta a_j, o ile pomiędzy tymi agentami zaufanie jest określone, podzielona przez iloczyn maksymalnej wartości zaufania i liczby takich par agentów. Średnie zaufanie do agenta a_j w kontekście c_k w chwili m_l jest wyrażone wzorem: $$t_{A \to a_j}^{c_k;m_l} = \frac{\sum_{i=1}^{n} t_{a_i \to a_j}^{c_k;m_l}}{t_{max} * \sum_{i=1}^{n} 1}$$ przy czym i jest takie, że: $a_i \in A$, $i \neq j$ oraz $\exists f_{trust}(a_i, a_j, c_k, m_l)$ – czyli że zaufanie agenta a_i do a_j jest określone.	Definicja 5.1.4., R5.1.4
liczba interakcji agentów rzetelnych z agentami rzetelnymi	to suma liczby interakcji, w której zarówno usługodawca a_j oraz usługobiorca a_i byli rzetelni: $$l_I^{A_B,A_B} = \sum_{i,j:a_i,a_j \in A_B, i \neq j} l_I^{R:a_i,P:a_j}$$	Definicja 5.1.5.1., R5.1.5
liczba interakcji agentów rzetelnych z agentami złośliwymi	suma liczby interakcji pomiędzy agentami, w której jeden z agentów był rzetelny, a drugi złośliwy: $$l_I^{A_B,A_M} = \sum_{i:a_i \in A_B, j:a_j \in A_M} l_I^{R:a_i,P:a_j} + \sum_{i:a_i \in A_M, j:a_j \in A_B} l_I^{R:a_i,P:a_j}$$	Definicja 5.1.5.2., R5.1.5
średni rzeczywisty wynik usług świadczonych przez agenta $\boldsymbol{a_j}$,	Średni rzeczywisty wynik usług świadczonych przez agenta $\boldsymbol{a_j}$, który wyświadczył przynajmniej jedną usługę, to iloraz sumy rzeczywistych wyników interakcji i liczby interakcji, w których ten agent był usługodawcą: $$\forall_{a_j \in A_P, l_I^{P:a_j} > 0} \; o_E^{P:a_j} = \frac{o_\Sigma^{P:a_j}}{l_I^{P:a_j}}$$	Definicja 5.1.5.3., R5.1.5
średnia liczba interakcji, w których agenci byli usługodawcami $\overline{\boldsymbol{l_I^{P:A}}}$,	Średnia liczba interakcji, w których agenci byli usługodawcami $\overline{\boldsymbol{l_I^{P:A}}}$, to iloraz sumy liczba interakcji, w których agenci, którzy wyświadczyli przynajmniej jedną usługę byli usługodawcami i liczby takich agentów: $$\overline{l_I^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} l_I^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$$	Definicja 5.1.5.4., R5.1.5,
średnia rzeczywista jakość usług ($\overline{\boldsymbol{o_E^{P:A}}}$)	Średnia rzeczywista jakość usług ($\overline{\boldsymbol{o_E^{P:A}}}$) jest średnią arytmetyczną średnich rzeczywistych wyników usług świadczonych przez agentów, którzy dostarczyli przynajmniej jedną usługę. $$\overline{o_E^{P:A}} = \frac{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} o_E^{P:a_j}}{\sum_{a_j \in A_P, l_I^{P:a_j} > 0} 1}$$	Definicja 5.1.5.5., R5.1.5,

ZAŁĄCZNIK 4 – OPIS ZNANYCH ATAKÓW NA SYSTEMY TRM

Dla celów niniejszej rozprawy zostało stworzone zestawienie ataków, na podstawie wcześniejszej pracy autora rozprawy [106] można stwierdzić, że: *„istotne zastrzeżenia dotyczące tego zestawienia są następujące:*

- *nie wymieniono wszystkich ataków zidentyfikowanych w literaturze w przypadku gdy różnice w specyfice ataków opisanych przez autorów były nieznaczące,*
- *niektóre z wyszczególnionych ataków funkcjonują w literaturze także pod innymi nazwami (przytoczono tylko te najczęściej stosowane),*
- *autor zdecydował się na pominięcie opisów ataków, przytaczając krótko jedynie jego ideę, umieszczono natomiast referencje do szczegółowych opisów w literaturze,*
- *nie wskazano referencji do wszystkich pozycji literaturowych, które opisują dany atak lub powołują się na niego (ze względu na to, że byłoby ich zbyt wiele), a jedynie do tych, które w sposób najbardziej kompleksowy i zarazem przejrzysty opisują dany atak."*

Zestawienie ataków zidentyfikowanych w literaturze i opracowanych przez autora zaprezentowano w tabeli.

Atak	**Przykłady publikacji**	**Skrócony opis ataku**	**Opis ataku (opisy ataków w znacznej mierze oparto na wcześniejszej publikacji autora rozprawy -** [85]**)**
wychwalanie (false-praise)	[1], [9], [69]	atak polega na zawyżaniu ocen innego agenta	*Atak ten polega na zawyżaniu przez złośliwego agenta rekomendacji na temat innego agenta. Wychwalanie zaburza spójność pomiędzy rzeczywistą (obserwowaną) rzetelnością agenta, a krążącymi opiniami o nim. Najprostszym i często wskazywanym w literaturze [1] sposobem obrony przed tym atakiem jest przechowywanie i monitorowanie przez agenty dwóch miar zaufania: jedną z nich jest zaufanie do świadczenia usług, a drugą oddzielną, jest zaufanie do wydawanych rekomendacji* [112].
oczernianie (bad-mouthing)	[1], [9], [69]	atak polega na zaniżaniu ocen innego agenta	*Jest to atak będący przeciwieństwem wychwalania. W celu jego przeprowadzenia złośliwy agent wydaje fałszywe, zaniżone rekomendacje o innych agentach. Dzięki temu atakujący zniechęca agenty do podejmowania współpracy z innymi rzetelnymi agentami, co pośrednio może przyczynić się do ograniczenia efektywności środowiska. Obrona przed tego typu atakiem jest taka sama jak w przypadku wychwalania i polega ona na oddzieleniu zaufania dotyczącego świadczonych usług od zaufania rekomendacyjnego* [112].

kreacja wielu tożsamości (Sybil attack)	[1], [9], [69]	atak polega na pozorowaniu istnienia większej liczby agentów, w celu zwiększenia istotności wydawanych rekomendacji	*Atak ten polega na tym, że pojedynczy agent zachowuje się w istocie tak, jakby stanowił grupę wielu agentów. Sama kreacja wielu tożsamości nie jest istotnym zagrożeniem, ale w połączeniu z innymi atakami, już takie zagrożenie stanowi. Atak tego typu wykorzystuje fakt, że agenty przy ocenie rzetelności innego agenta będą kierowały się opiniami innych agentów, stosując zasadę większościową – im więcej agentów będzie wyrażało daną opinię tym większe prawdopodobieństwo, że dany agent będzie się nią kierować. Jak jest podkreślane w literaturze [1] obrona przed atakami fałszowania tożsamości nie leży w gestii systemu TRM ale w mechanizmach identyfikacji i uwierzytelnienia stosowanym w środowisku. Najprostszym sposobem obrony przed nim jest poniesienie przez dany agent kosztu dołączenia do sieci (np. poprzez wykonanie złożonej obliczeniowo operacji) albo też centralne zarządzanie tożsamościami* [74].
kreacja nowej tożsamości (whtewashing, new-comer)	[1], [9], [69]	atak polega na pozorowaniu, że skompromitowany agent dopiero pojawił się w środowisku	*Wykonując ten atak, agent rezygnuje ze swojej dotychczasowej tożsamości i kreuje nową. Atak jest wykorzystywany w sytuacji gdy dany atakujący, wskutek swojego nierzetelnego zachowania w stosunku do innych agentów, sprawił, że inne agenty obniżyły do niego zaufanie. Po doprowadzeniu do takiej sytuacji może on więc przedstawiać się innym agentom jako całkiem nowy byt (np. posługując się innym identyfikatorem), dzięki temu będzie on posiadał zaufanie ze strony innych agentów na poziomie takim jak agenty pojawiające się w środowisku (które jest zwykle wyższe niż agentów skompromitowanych). Obrona przed tym atakiem może być podobna do obrony przed atakiem wielu-tożsamości.*
atak stały (constant)	[1], [9], [69]	atak polega na ciągłym świadczeniu nierzetelnych usług	*Atak polega na ciągłym świadczeniu usług o minimalnej jakości.*
oscylacja zachowania (On-off attack)	[1], [9], [69]	atak polega na cyklicznie powtarzających się okresach świadczenia rzetelnych i nierzetelnych usług	*Atak ten polega na tym, że agent atakujący przez pewien czas zachowuje się w stosunku do innych agentów rzetelnie (dzięki czemu buduje wysokie zaufanie do siebie w opinii innych agentów), po to by następnie zachowywać się w sposób złośliwy (inne agenty chętnie podejmują współpracę z tym agentem z racji na wysokie zaufanie do niego, ale doświadczają jego nierzetelności). Następnie ten schemat się powtarza. Najprostszym sposobem obrony przed tego typu atakiem jest przywiązywanie większej wagi do nierzetelnych działań agenta niż do rzetelnych. Odmianą takiego sposobu postępowania jest użycie adaptacyjnego współczynnika zapominania (adaptive forgetting factor* [1], [112]*), dzięki czemu zaufanie jest wyznaczane na podstawie pewnej liczby ostatnio zaobserwowanych zachowań agenta, a wpływ ostatniej obserwacji na zaufanie do agenta jest większy niż wcześniejszych interakcji.*
niespójne zachowanie (conflicting bahaviour)	[1], [9], [69]	atak polega na rzetelnym świadczeniu usług dla pewnej grupy rzetelnych agentów, a jednocześnie nierzetelnym świadczeniu usług dla innych rzetelnych agentów	*Atak ten polega na zachowywaniu się odmiennie w stosunku do różnych agentów i dzięki temu może skutkować obniżeniem zaufania do rekomendacji innych agentów. W ataku tym, agent zachowuje się zawsze rzetelnie w stosunku do jednej grupy rzetelnej agentów, natomiast zachowuje się zawsze nierzetelnie w stosunku do innej grupy agentów rzetelnych. Agenty w obu grupach wymieniają się między sobą rekomendacjami na temat agenta atakującego. Grupa w stosunku, do której agent atakujący zachowuje się rzetelnie, otrzymuje od drugiej grupy niskie rekomendacje na temat tego agenta, co jest sprzeczne z ich własnymi obserwacjami, wskutek tego agenty z tej grupy obniżają zaufanie do agentów z drugiej grupy. Analogicznie sytuacja wygląda w drugiej grupie. Dzięki temu agent atakujący osiągnął ograniczenie zaufania pomiędzy tymi dwiema grupami agentów rzetelnych, co negatywnie przełoży się na efektywność środowiska.*

slander attack	[73]	kombinacja ataków oscylacji zachowania oraz wychwalania o specyficznym przebiegu – skuteczny tylko przeciw niewielkiej liczbie systemów TRM	*W artykule* [73] *rozważany jest bardzo specyficzny przypadek zastosowania systemu TRM: agenty określają zaufanie tylko do agenta sąsiedniego, a korzystają przy tym z rekomendacji dostarczonych także tylko przez agenty sąsiednie. Atak ten zakłada scenariusz w którym agent A jest obserwowany przez agenta B. Agent A zachowywał się rzetelnie przez pewien czas i w związku z tym miał wysokie zaufanie. Agent ten jednakże zmienia swoje zachowanie, a dodatkowo pewna część agentów sąsiadujących z agentami A i B zaczyna prezentować agentowi B nieprawdziwe rekomendacje i nadal przekonywać go, że agent A zachowuje się prawidłowo. W przywoływanym artykule sposobem obrony przed tego typu atakami jest porównywanie rekomendacji od wielu agentów w celu identyfikacji agentów złośliwych. Atak ten ma zastosowanie jedynie do wąskiej grupy systemów TRM.*
złośliwy kolektyw	[74]	atak polega na złośliwym zachowaniu grupy agentów i jednocześnie zawyżaniu reputacji innych złośliwych agentów	*W ataku tym, złośliwe agenty zawsze zachowują się w sposób nierzetelny, ale prezentują wysokie rekomendacje w stosunku do innych złośliwych agentów. Jest to atak stały i wychwalania wykonywany przez wielu agentów złośliwych. Proponowanym sposobem obrony przed takim atakiem jest ocena trafności rekomendacji wydawanych przez agenty i obniżanie zaufania do agenta dostarczającego nieprawidłowych rekomendacji.*
złośliwy kolektyw z kamuflażem	[74]	atak polega na oscylacji zachowania złośliwych agentów i jednocześnie zawyżaniu reputacji innych agentów	*Atak ten jest podobny do ataku „złośliwy kolektyw". Różnicą jest to, że w tym ataku złośliwe agenty nie zachowują się nierzetelnie przez cały czas. Jest to rozbudowanie poprzedniego ataku o użycie ataku oscylacji zachowania. Obroną przed atakiem tego typu może być wykorzystanie oddzielnych miar zaufania do świadczenia usług i rekomendacji oraz karanie za zmienne zachowanie.*
złośliwi szpiedzy	[74]	w ataku występują dwie grupy złośliwych agentów – jedna stosuje atak oscylacji zachowania i wychwalanie, a druga tylko atak wychwalania	*W ataku istnieją dwa typy (grupy) złośliwych agentów: pierwszy typ to agenty zachowujące się nierzetelnie i dające pozytywne rekomendacje o wszystkich innych złośliwych agentach, drugi typ to agenty zachowujące się zawsze rzetelnie ale dające pozytywne rekomendacje o wszystkich innych złośliwych agentach. Jest to rozbudowany wariant poprzedniego ataku. Obroną, podobnie jak w poprzednim przypadku mogą być oddzielne miary zaufania do świadczenia usług i rekomendacji oraz karanie za zmienne zachowanie.*
obniżenie reputacji	[74]	kombinacja ataków złośliwy kolektyw i oczerniania w wersji kooperacyjnej	*Atak jest podobny do ataku „złośliwy kolektyw", z tym, że dodatkowo złośliwe agenty oczerniają rzetelne agenty. Obrona, tak jak w przypadku poprzednich ataków, polega na wykorzystaniu oddzielnych miar zaufania do świadczenia usług i rekomendacji oraz karanie za zmienne zachowanie.*
częściowo złośliwy kolektyw	[74]	tak jak w ataku złośliwy kolektyw, z tym że złośliwe agenty zachowują się nierzetelnie tylko w przypadku ściśle określonego typu interakcji	*W ataku tym agenty zachowują się nierzetelnie jako dostawcy określonej usługi, natomiast rzetelnie jako dostawcy innych usług, dodatkowo wydają zawyżone rekomendacje o innych złośliwych agentach. Obroną przed tego typu atakiem jest zdefiniowanie i monitorowanie oddzielnych miar zaufania dla poszczególnych usług.*

grupowy atak oscylacyjny (oscillation attack)	[10], [75]	w ataku uczestniczą dwie grupy złośliwych agentów, jedna z nich wykonuje atak oscylacji zachowania, a druga dostarcza prawidłowych rekomendacji, ale tylko dla części istotnych agentów	*Atak ten jest oparty na ataku oscylacji zachowania. W grupowym ataku oscylacyjnym w jednym czasie pewna grupa atakujących agentów dostarcza nieprawidłowe rekomendacje dla agentów będących celem ataku, natomiast inna grupa atakujących stara się zwiększyć swoje zaufanie poprzez dostarczanie prawidłowych rekomendacji, jednakże zachowuje się rzetelnie tylko w stosunku do tych agentów, które nie są szczególnie istotne (znajdują się poza celem ataku). Po pewnym czasie, te dwie grupy atakujących agentów zamieniają się rolami. Podobnie jak w przypadku ataków polegających na oscylacji zachowania, dokonywanych przez pojedyncze agenty, tak i w przypadku tego ataku podstawowym sposobem obrony jest przywiązywanie większej wagi do nierzetelnych zachowań agentów niż do zachowań rzetelnych.*
RepTrap	[10], [76]	złośliwe agenty stosują atak oczerniania w stosunku do kilku agentów z najwyższą reputacją, ale mających niewiele interakcji z innymi agentami	*W ataku RepTrap złośliwe agenty dokonują analizy interakcji zachodzących między rzetelnymi agentami i wybierają jako cel ataku te spośród rzetelnych agentów, które cieszą się wysokim zaufaniem, jednocześnie mając niewiele interakcji z innymi agentami. Wtedy złośliwe agenty przeprowadzają atak oczerniania dostarczając wszystkim agentom negatywnych rekomendacji na temat wybranych agentów. Jeżeli liczba atakujących jest znacznie większa niż liczba rzetelnych agentów mających interakcję z obwinianymi agentami, to pozostałe agenty zaczynają traktować rekomendacje atakujących jako prawidłowe, natomiast rekomendacje dostarczane przez rzetelne agenty jako potencjalnie złośliwe. W związku z tym agenty zwiększają zaufanie do złośliwych agentów, natomiast obniżają zaufanie zarówno do agentów obwinianych jak i agentów rzetelnych mających bezpośrednią interakcję z agentami obwinianymi (i prezentującymi wysokie rekomendacje o tych agentach). Warto zauważyć, że atak ten zawiera w sobie elementy ataków polegających na oczernianiu i ataku niespójnego zachowania, ale jego zasadniczym elementem jest sposób selekcji celu ataku (agenty o wysokim zaufaniu, niemające zbyt wielu interakcji, oraz agenty z nimi współpracujące).*
wyrocznia	[45], [85]	w ataku uczestniczą dwie grupy złośliwych agentów aktywnie ze sobą współpracujący ch	*W ataku tym zdefiniowano dwie grupy złośliwych agentów. W pierwszej grupie znajdują się agenty (nazywane agentami obserwowanymi), których zadaniem jest generalnie rzetelne wykonywanie żądań innych agentów w sieci. Jednak od czasu do czasu agenty te zachowują się w sposób nierzetelny (stosują więc atak oscylacji zachowania). W drugiej grupie znajdują się agenty (zwane wyroczniami), które zawsze „przewidują" w jaki sposób obserwowane agenty się zachowają świadcząc usługę. Agenty te (wyrocznie) dzielą się swoimi rekomendacjami z innymi agentami. Obydwie grupy agentów uzgadniają w jaki sposób zachowają się agenty obserwowane, a wobec tego rekomendacje wyroczni są zawsze słuszne. Natomiast pozostałe agenty wydające rekomendacje o obserwowanych agentach będą dawały rekomendacje pozytywne (z racji tego, że agenty obserwowane zwykle zachowują się rzetelnie), będą więc myliły się w momencie gdy obserwowane agenty zachowają się nierzetelnie. Wskutek takiego działania dokonana a posteriori ocena trafności rekomendacji agentów spowoduje, że zaufanie do agentów-wyroczni wzrośnie, natomiast zaufanie do pozostałych agentów wydających rekomendacje, zostanie obniżone.*

ZAŁĄCZNIK 5 – SPECYFIKACJA TECHNICZNA NARZĘDZIA TRM-RET

Zależności – używane biblioteki

Narzędzie zaimplementowano w języku python 3.6. Do funkcjonowania narzędzia konieczne jest użycie następujących wersji pakietów (bibliotek zewnętrznych):

- document==1.0
- docx==0.2.4
- kaleido==0.2.1
- lxml==4.9.2
- numpy==1.23.5
- packaging==22.0
- pandas==1.5.2
- patsy==0.5.3
- Pillow==9.4.0
- plotly==5.11.0
- plotly-express==0.4.1
- python-dateutil==2.8.2
- python-docx==0.8.11
- pytz==2022.6
- scipy==1.9.3
- six==1.16.0
- statsmodels==0.13.5
- tenacity==8.1.0

Wymagania sprzętowe

Narzędzie zostało z powodzeniem używane na maszynie wirtualnej o niewygórowanych parametrach (wymienionych poniżej), w przypadku lepszych parametrów sprzętowych, na których będzie uruchomione narzędzie, przeprowadzane badania będą przebiegały szybciej[78].

Parametry maszyny wirtualnej na której były przeprowadzane badania:

- 4 GB RAM
- Procesor Intel Core 2.20 GHz
- Dysk 100 GB
- System operacyjny: Linux Debian

[78] Pojedyncze badanie spośród tych wymienionych w rozprawie nigdy nie trwało dłużej niż kilkadziesiąt sekund (nie dotyczy to metody MEAEM, dla której badania trwały znacznie dłużej – w zależności od wariantu – od kilku minut do kilku godzin).

SPIS RYSUNKÓW

SPIS TABEL