МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА

На правах рукописи

Терёхина Ирина Юрьевна

Методы выявления аномалий в условиях смеси технологических процессов, сопровождающих наблюдаемый объект

Специальность 2.3.6 —
«Методы и системы защиты информации, информационная безопасность»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2024

Диссертация подготовлена на кафедре информационной безопасности факультета Вычислительной математики и кибернетики МГУ имени М. В. Ломоносова.

Научный руководитель: **Грушо Александр Александрович,** доктор физико-математических наук, профессор.

Официальные оппоненты: **Бурдонов Игорь Борисович,** доктор физико-математических наук, доцент, Институт системного программирования РАН имени В. П. Иванникова, главный научный сотрудник.

Сергеев Игорь Сергеевич, доктор физико-математических наук, ФГУП «Научно-исследовательский институт "Квант"», начальник лаборатории.

Шелупанов Александр Александрович, доктор технических наук, профессор, ФГБОУ ВО "Томский государственный университет систем управления и радиоэлектроники", президент.

Защита диссертации состоится 30 октября 2024 г. в 16 часов 45 минут на заседании диссертационного совета МГУ.012.3 Московского государственного университета имени М.В.Ломоносова по адресу: Российская Федерация, 119234, Москва, ГСП-1, Ленинские горы, д. 1, Главное здание МГУ, Механико-математический факультет, комн. 14-08.

Email: vasenin@msu.ru.

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27), а также на портале: https://dissovet.msu.ru/dissertation/3095.

Автореферат разослан 24 сентября 2024 года.

Ученый секретарь
диссертационного совета МГУ.012.3,
канд. физ.-мат. наук Галатенко А. В.

Общая характеристика работы

Актуальность темы

Диссертация посвящена исследованию подходов к решению задачи поиска отклонений от некоторого принятого режима функционирования (далее аномалий) реальных объектов и процессов их сопровождения. Под процессом в контексте настоящей работы понимается описание информационной технологии, как множество конечных последовательностей атомарных инструкций, предназначенное для сопровождения объектов и/или услуг для пользователя (создание, контроль состояния и управление).

В диссертации рассматриваются объекты, процессы сопровождения которых имеют техническую природу и могут быть формально описаны. Также предполагается, что рассматриваемые объекты могут быть описаны в рамках математических моделей, используемых в контексте диссертации, и вопросы обеспечения их безопасности от деструктивных воздействий, а, значит, вопрос поиска аномалий имеет практическое значение.

В концептуальном плане модели поиска аномалий, представленные в диссертации, разрабатываются в контексте решения двух задач. Первая из них направлена на построение математической модели процесса по доступным логам (стереотипным режимам) функционирования процесса. Вторая обеспечивает поиск аномалий с использованием построенной математической модели процесса, которая описывает легальное (принятое, традиционное) его поведение.

Поиск аномалий процесса с использованием восстановленной по его логам модели является актуальным вопросом с позиции требований обеспечения информационной безопасности процесса по перечисленным далее причинам.

1. Необходимость обеспечения информационной безопасности возникает в большом количестве прикладных областей — от многопользовательских систем, где задача поиска аномалий может возникать в рамках обеспечения целостности данных и разграничения доступа к ним; до областей, где происходит сбор и систематизация большого количества отчетов (логов), с последующим поиском подходов к выявлению и обнаружению вторжений.
2. Если есть алгоритм построения модели процесса по его логу, и полученная модель отражает возможные варианты исполнения процесса, то этот факт означает, что найден прототип этого процесса без дополнительной информации о том, как соответствующая процессу информационная технология устроена “внутри”. Таким образом, информационная технология может стать более интерпретируемой и понятной для восприятия человеком.
3. В информационной безопасности достаточно сложно дать формальное определение того, что именно можно назвать “аномалией”.

Автор диссертации исходит из следующего интуитивного предположения: если известна модель "правильного" функционирования системы или процесса, то определение аномального поведения для системы или процесса может быть дано следующим образом — это такое поведение, которое не может быть порождено этой моделью. В этом предположении дано, как и определение того, что можно назвать аномалией, так и понятен подход к построению алгоритма, осуществляющего поиск аномалий. Такое определение аномалий означает, что при известной модели процесса, задача поиска аномалий может быть переформулирована в задачу соответствия некоторого исполнения процесса построенной модели, решение которой, как правило, является менее трудоемким.

Вопрос построения математической модели по данным некоторой функционирующей информационной технологии (процесса, рабочего процесса) часто формулируется как задача построения модели (обнаружения) процесса. В настоящее время опубликовано и продолжает появляться большое число исследований, посвященных методам восстановления технологических процессов, реализующих информационные технологии, по логам и другим наблюдаемым последовательностям действий, сопровождающих выполнение информационных технологий[1]. Впервые подходы к решению этой задачи в контексте технологических процессов представлены в[2], после чего появлялось большое число других подходов к ее решению[3]. Традиционно вводятся ряд понятий, которые будут использоваться и в настоящей диссертации. Понятие *одного исполнения процесса* — наблюдение за изменением атрибутов процесса в ограниченных временных рамках. Исполнению

[1] *Leno, V.* Robotic process mining: vision and challenges / V. Leno, A. Polyvyanyy, M. Dumas, M. La Rosa, F. Maggi // Business & Information Systems Engineering. 2021. Т. 63, № 3. С. 301—314; *Andrews, R.* Quality-informed semi-automated event log generation for process mining / R. Andrews, C. G. van Dun, M. T. Wynn, W. Kratsch, M. Röglinger, A. H. ter Hofstede // Decision Support Systems. 2020. Т. 132. С. 113265; *Pham, D.-L.* Process-Aware Enterprise Social Network Prediction and Experiment Using LSTM Neural Network Models / D.-L. Pham, H. Ahn, K.-S. Kim, K. P. Kim // IEEE Access. 2021. Т. 9. С. 57922—57940; *Shakya, S.* Process mining error detection for securing the IoT system / S. Shakya // Journal of ISMAC. 2020. Т. 2, № 03. С. 147—153; *Cerezo, R.* Process mining for self-regulated learning assessment in e-learning / R. Cerezo, A. Bogarin, M. Esteban, C. Romero // Journal of Computing in Higher Education. 2020. Т. 32, № 1. С. 74—88.

[2] *Agrawal, R.* Mining process models from workflow logs / R. Agrawal, D. Gunopulos, F. Leymann // International Conference on Extending Database Technology. Springer. 1998. С. 467—483.

[3] *Cook, J. E.* Discovering models of software processes from event-based data / J. E. Cook, A. L. Wolf // ACM Transactions on Software Engineering and Methodology (TOSEM). 1998. Т. 7, № 3. С. 215—249; *Mannila, H.* Discovery of frequent episodes in event sequences / H. Mannila, H. Toivonen, A. I. Verkamo // Data mining and knowledge discovery. 1997. Т. 1, № 3. С. 259—289; *Schimm, G.* Process miner—a tool for mining process schemes from event-based data / G. Schimm // European Workshop on Logics in Artificial Intelligence. Springer. 2002. С. 525—528; *Herbst, J.* Ein induktiver ansatz zur akquisition und adaption von workflow-modellen / J. Herbst. Tenea Verlag Ltd., 2004.

процесса ставится в соответствие понятие *трассы* — непустого слова в некотором конечном непустом алфавите возможных *действий* процесса, наблюдений за изменением конечного подмножества атрибутов, упорядоченных по времени. Восстановленная модель процесса напрямую зависит от заданного *лога* (лога исполнений процесса, лога событий) — конечного неупорядоченного множества трасс, которое предполагается достаточным для описания функционирования процесса.

Для одного процесса, сопровождающего функционирование некоторого объекта, могут быть построены различные математические модели, описывающие этот процесс. Также выбор подходящей математической модели может зависеть от того, поиск решений каких именно задач относительно данного процесса необходимо осуществить. Для задач обеспечения информационной безопасности можно привести следующие примеры того, что может задавать модель процесса.

- Модификация ресурсов субъектами, влияющими на реализацию информационной технологии. В этом случае исследование подходов к решению задачи поиска аномалий сводится к поиску неправомерного использования ресурсов, то есть к построению методов решения задачи управления/разграничения доступа к информационным ресурсам.
- Модификация данных во времени по мере реализации информационной технологии. Тогда вопрос поиска аномалий может быть сведен к вопросу обеспечения целостности данных.

Цели и задачи

Целью диссертационной работы является разработка моделей и алгоритмов для решения задачи поиска аномалий в условиях функционирования нескольких процессов.

Для достижения поставленной цели необходимо решение следующих задач.

1. Поиск эффективных методов моделирования процесса по некоторому логу и методов выявления аномалий, с использованием известной модели процесса.
2. Разработка эффективного по времени выполнения относительно проверяемой на аномальность трассы метода обнаружения аномалий с помощью различных типов математических моделей процесса.
3. Разработка метода построения моделей нескольких процессов в зависимости от свойств лога, описывающего исполнения процессов.
4. Получение оценок сложности алгоритмов решения задачи построения моделей процесса и алгоритмов обнаружения аномалий с помощью моделей процесса, проведение экспериментов.

5. Обобщение алгоритмов решения задачи поиска аномалий на случай, когда проверяемая трасса содержит в себе данные нескольких процессов.

Основные положения, выносимые на защиту

На защиту выносятся обоснование актуальности решаемой задачи, методология, принятая для исследования, научная новизна, теоретическая и практическая значимости работы, а также следующие положения, которые подтверждаются результатами исследования, представленными в Заключении диссертации.

1. Результаты исследования возможности использования математических моделей в виде сетей Петри для решения задачи построения модели процесса и для решения задачи поиска аномалий.
2. Результаты исследования возможности использования математических моделей в виде ациклических ориентированных графов (DAG) для решения задачи построения модели процесса и для решения задачи поиска аномалий.
3. Алгоритмы и соответствующие им оценки сложности выполнения по времени для решения задачи построения моделей нескольких процессов в виде ациклических ориентированных графов.
4. Алгоритмы и соответствующие им оценки сложности выполнения для решения задачи выявления аномалий в некоторой трассе при использовании моделей процессов в виде ациклических ориентированных графов. Обобщение предложенных алгоритмов и оценка сложности выполнения для случая, когда в проверяемой на аномальность трассе содержатся данные нескольких процессов.

Научная новизна

Предложено решение задачи поиска аномалий с помощью математических моделей, описывающих представления нескольких одновременно функционирующих технологических процессов.

Предложены решения и получены оценки временной сложности восстановления множества экземпляров информационных технологий в форме ациклических ориентированных графов. Используя построенные модели процесса в виде ациклических ориентированных графов, предложено решение для задачи поиска аномалий. Предложено обобщение методов восстановления моделей процессов в случаях функционирования нескольких информационных технологий.

Получены результаты, демонстрирующие отсутствие возможности однозначного восстановления модели процесса, представленной в терминах простейших сетей Петри. Этот факт исключает возможность применения данного математического аппарата для решения задачи поиска аномалий.

Разработан и описан подход, позволяющий выбирать математические модели для решения конкретных задач информационной безопасности на примере задачи поиска аномалий. Показано, что критерий простоты описания математической модели не может быть решающим в выборе подходящей модели.

Все результаты, полученные в диссертации, являются новыми.

Теоретическая и практическая значимость

Основная теоретическая значимость исследования состоит в том, что автор предлагает новые решения задач распознавания и построения процессов, описывающих функционирование информационных технологий, используя аппарат глубокого анализа эквивалентных представлений этих процессов.

Практическая значимость полученных результатов состоит в возможности применения результатов в анализе защищенности реальных систем.

Методология и методы исследования

В работе используются методы дискретной математики, теории графов, теории сложности управляющих систем.

Соответствие паспорту научной специальности

Полученные в диссертации результаты соответствуют паспорту специальности 2.3.6 — методы и системы защиты информации, информационная безопасность (физико-математические науки).

Диссертация представляет результаты исследований в области информационной безопасности. В работе используются методы дискретной математики, теории графов, теории сложности управляющих систем.

Области исследования:

3. Методы, модели и средства выявления, идентификации и классификации угроз нарушения информационной безопасности объектов различного вида и класса.
15. Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности.
16. Модели, методы и средства обеспечения внутреннего аудита и мониторинга состояния объекта, находящегося под воздействием угроз нарушения его информационной безопасности.

Достоверность

Достоверность полученных результатов обеспечивается приведенными точными математическими доказательствами. Работы других авторов,

используемые в диссертации, отмечены соответствующими ссылками. Результаты диссертации опубликованы в открытой печати.

Апробация работы

Результаты диссертации докладывались на следующих конференциях и научно-исследовательских семинарах.

1. Семинар «Компьютерная безопасность» под руководством д.ф.-м.н., проф. Грушо А.А., д.т.н., проф. Тимониной Е.Е, кафедра информационной безопасности факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, неоднократно с 2017 по 2021 год.
2. Семинар «Компьютерная безопасность» под руководством к.ф.-м.н. Галатенко А.В., к.ф.-м.н. Александрова Д.Е., кафедра МаТИС механико-математического факультета МГУ им. М.В. Ломоносова, 2022 год.
3. Кафедральный семинар кафедры информационной безопасности под руководством д.т.н., акад. РАН Соколова И.А., факультет вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, март 2022 года.
4. Четырнадцатый международный семинар «Дискретная математика и ее приложения» имени академика О. Б. Лупанова, Москва, МГУ им. М.В. Ломоносова, 20–24 июня 2022 год.
5. Семинар «Проблемы современных информационно-вычислительных систем» под руководством д.ф.-м.н., проф. В.А. Васенина, механико-математический факультет МГУ им. М.В. Ломоносова, 7 марта 2023 год.
6. Научная конференция «Ломоносовские чтения 2023», Москва, МГУ им. М.В. Ломоносова. Секция «Вычислительной математики и кибернетики», 4-14 апреля 2023 год.

Личный вклад

Автор сформулировал и доказал представленные в диссертационной работе утверждения, теоремы, направленные на решение задачи поиска аномалий и решение задачи построения формальных моделей процессов. Все результаты получены автором самостоятельно.

Публикации

Результаты, выносимые на защиту, изложены в 4 статьях ([1—3; 6]), 3 из которых опубликованы в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ имени М. В. Ломоносова по специальности 2.3.6, а также в 2 сборниках трудов конференций ([4; 5]).

Объем и структура работы

Диссертация состоит из введения, трёх глав, заключения. Полный объём диссертации составляет 122 страницы, включая 43 рисунка и 2 таблицы. Список литературы содержит 73 наименований.

Краткое содержание работы

В **Введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируются цели и задачи работы, излагается научная новизна.

Первая глава посвящена обзору существующих результатов для решения задачи построения модели процесса и задачи поиска аномалий.

Вторая глава посвящена использованию математических моделей в виде сетей Петри для решения задачи поиска аномалий. Основные результаты главы представлены в публикациях [1; 2].

Далее автором вводится ряд определений, необходимых для формулировок условий корректности, которым должна удовлетворять модель. Мотивация выбора именно таких условий корректности согласуется с работами, которые посвящены исследованиям о выборе наилучшей модели для описания функционирующей информационной технологии[4]. Определения вводятся по аналогии с работой[5], ряд определений были скорректированы в сторону упрощения.

Определение (Сеть Петри)**.** Сеть Петри N — тройка (P, T, F), где:

1. P — конечное множество, *места*;
2. T — конечное множество, *переходы*, $P \cap T = \varnothing$;
3. $F \subseteq (P \times T) \cup (T \times P)$ — множество направленных некратных дуг, отношение инцидентности.

Для описания динамического поведения сети Петри вводится определение разметки над множеством мест P. В рамках настоящей работы рассматриваются разметки, в которых в каждом месте сети N может находиться не более одной метки.

Определение (Размеченная сеть Петри)**.** Пусть $N = (P, T, F)$ — сеть Петри. Разметка s — множество мест из P, в которых есть метка. Пара $(N = (P, T, F), s)$ — размеченная сеть Петри.

Определение (Входной и выходной узел)**.** Пусть $N = (P, T, F)$ — сеть Петри. Элементы из $P \cup T$ — *узлы*. Пусть $x, y \in P \cup T$.

[4] *Mendling, J.* Seven process modeling guidelines (7PMG) / J. Mendling, H. A. Reijers, W. M. van der Aalst // Information and Software Technology. 2010. Т. 52, № 2. С. 127—136.

[5] *Van der Aalst, W.* Workflow mining: Discovering process models from event logs / W. Van der Aalst, T. Weijters, L. Maruster // IEEE transactions on knowledge and data engineering. 2004. Т. 16, № 9. С. 1128—1142.

Узел x — *входной узел* узла y, если существует направленная дуга из x в y, то есть $(x, y) \in F$. Входной узел узла y обозначается следующим образом:

$$\bullet y = \{x | (x, y) \in F\}.$$

Узел y — *выходной узел* узла x, если $(x, y) \in F$. Выходной узел узла x обозначается следующим образом:

$$x\bullet = \{y | (x, y) \in F\}.$$

Динамическое поведение сети Петри задается следующим образом.

Определение (Правило срабатывания переходов)**.** Пусть $(N = (P, T, F), s)$ — размеченная PT-сеть. Переход $t \in T$ *активирован*, $(N, s)[t\rangle$, если $\bullet t \subseteq s$. Правило срабатывания переходов $\_[\_\rangle\_ \subseteq \mathcal{N} \times T \times \mathcal{N}$ — наименьшее отношение, удовлетворяющее: $\forall (N = (P, T, F), s) \in \mathcal{N}$ и $\forall t \in T$:

$$(N, s)[t\rangle \implies (N, s)[t\rangle(N, s \setminus (\bullet t) \cup (t\bullet))$$

Определение (Достижимые разметки)**.** Пусть (N, s_0) — размеченная сеть Петри из $\mathcal{N}$. Разметка s *достижима* из начальной разметки s_0, если существует последовательность активированных переходов, таких, что их последовательное срабатывание приведут из разметки s_0 в разметку s. Множество достижимых разметок (N, s_0) обозначается $[N, s_0\rangle$.

Определение (Срабатывающая последовательность)**.** Пусть (N, s_0), где $N = (P, T, F)$ — размеченная сеть Петри. Последовательность переходов $\sigma \in T^*$ называется *срабатывающей* для (N, s_0), если $\exists n \in \mathbb{N} \cup \{0\}$ такое, что существуют разметки $s_1, \ldots, s_n$ и переходы $t_1, \ldots, t_n \in T$ такие, что $\sigma = t_1 \ldots t_n$ и $\forall i, 0 \leq i < n$, $(N, s_i)[t_{i+1}\rangle$ и $s_{i+1} = s_i \setminus (\bullet t_{i+1}) \cup (t_{i+1}\bullet)$.

Последовательность σ называется *активированной* в разметке s_0: $(N, s_0)[\sigma\rangle$. Срабатывание последовательности σ ведет к разметке s_n: $(N, s_0)[\sigma\rangle(N, s_n)$

Если $n = 0$, то $\sigma = \varepsilon$, где ε – пустая последовательность. Таким образом, пустая последовательность также является срабатывающей последовательностью для (N, s_0).

Рассматриваются сети Петри с двумя выделенными местами: $i \in P$ —- входное место сети, $o \in P$, $i \neq o$, —- выходное место сети. Предполагается, что в сеть N новая метка может поступать только во входное место и покидать сеть через выходное место сети. Также далее рассматриваются только связные сети Петри (workflow nets[6]).

Пусть R^{-1} — обратное отношение к отношению R, а R^* — рефлексивное и транзитивное замыкание отношения R.

[6] *Van der Aalst, W.* Workflow mining: Discovering process models from event logs / W. Van der Aalst, T. Weijters, L. Maruster // IEEE transactions on knowledge and data engineering. 2004. Т. 16, № 9. С. 1128—1142.

Определение (Связная сеть Петри)**.** Сеть Петри $N = (P,T,F)$ *связна*, если для любых $x, y \in P \cup T$, справедливо $x(F \cup F^{-1})^* y$.

Пример: Размеченная сеть Петри $(N = (P,T,F), \{i\})$, где $P = \{i,o,p_1,p_2\}$, $T = \{t_1,t_2\}$,

$$F = \{(i,t_1), (t_1,p_1), (t_1,p_2), (p_1,t_2), (t_2,o)\},$$

Рис. 1. Метка в месте i активирует переход t_1. Множество входящих мест в переход t_1: $(\bullet t_1) = \{i\}$, множество выходящих мест $(t_1\bullet) = \{p_1,p_2\}$.

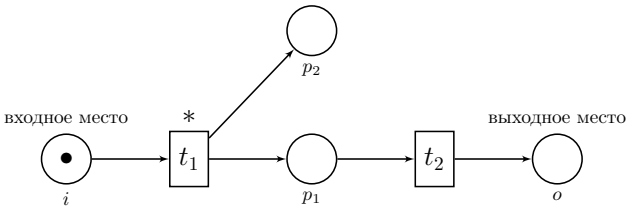

Рис. 1 — Размеченная сеть $(N,\{i\})$. Активированный переход отмечен $*$

Пусть срабатывает переход t_1. Если в сети Петри есть несколько активированных переходов, то в следующий момент времени может сработать любое подмножество из них[7]. В работе предполагается, что в один момент времени может сработать только 1 из активированных переходов. По правилу срабатывания переходов, новая разметка для сети N после срабатывания перехода t_1: $s \cup (t\bullet) \setminus (\bullet t) = \{i\} \cup \{p_1,p_2\} \setminus \{i\} = \{p_1,p_2\}$, становится активированным переход t_2, Рис. 2.

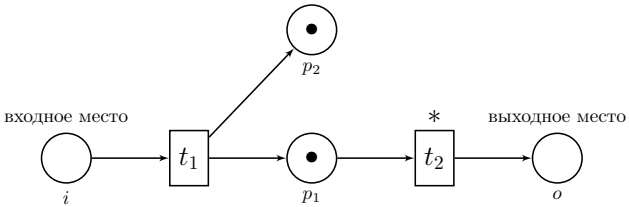

Рис. 2 — Размеченная сеть $(N,\{p_1,p_2\})$

Пусть срабатывает переход t_2. Множество входящих мест в переход t_2: $(\bullet t_2) = \{p_1\}$, множество выходящих мест $(t_2\bullet) = \{o\}$. Новая разметка после срабатывания перехода t_2 по правилу срабатывания переходов: $\{p_1,p_2\} \cup \{o\} \setminus \{p_1\} = \{p_2,o\}$, Рис. 3.

[7] *Питерсон, Д.* Теория сетей Петри и моделирование систем / Д. Питерсон. Мир, 1984.

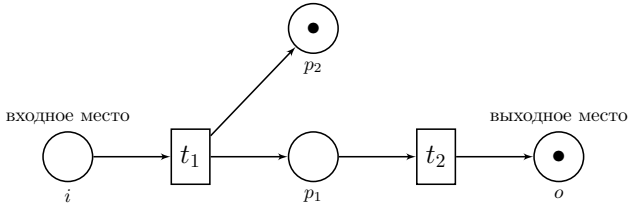

Рис. 3 — Размеченная сеть $(N,\{p_2,o\})$

Определение (Замкнутость и безопасность)**.** Размеченная сеть $(N = (P,T,F),s)$ *замкнута*, если множество достижимых разметок $[N,s\rangle$ конечно. Размеченная сеть *безопасна*, если $\forall s' \in [N,s\rangle$ и $\forall p \in P$, $s'(p) \leq 1$, где $s'(p)$ — количество меток в разметке s' в месте p.

Определение (Мертвые переходы, жизнеспособность)**.** Пусть $(N = (P,T,F),s)$ — размеченная PT-сеть. Переход $t \in T$ *мертвый*, если не существует достижимой разметки $s' \in [N,s\rangle$ такой, что $(N,s')[t\rangle$.

Сеть (N,s) — *жизнеспособна*, если для всех достижимых разметок $s' \in [N,s\rangle$, $\forall t \in T$, существует достижимая разметка $s'' \in [N,s'\rangle$ такая, что $(N,s'')[t\rangle$.

Первое условие корректности формулируется следующим образом.

Определение (Свойство надежности)**.** Пусть $N = (P,T,F)$ — сеть Петри с входным местом i и выходным местом o. N *надежна*, если справедливы следующие утверждения.

1. $(N,[i])$ безопасна.
2. $(N,[i])$ корректно завершается: $\forall s \in [N,[i]\rangle$, $o \in s \implies s = \{o\}$.
3. $(N,[i])$ не содержит мертвых переходов.

Можно видеть, что это условие корректности налагает минимальные интуитивные необходимые ограничения на сеть Петри для возможности дальнейшего поиска аномалий. Свойство безопасности отвечает за наличие не более одной метки в одном месте сети Петри в один момент времени; свойство корректного завершения говорит о том, что если метка попала в выходное место, эта метка должна быть единственной в сети Петри; свойство отсутствия мертвых переходов говорит о том, что в сети Петри не должно быть переходов, которые не активируются ни в одной из достижимых разметок.

Определение (Действие, трасса, лог)**.** Пусть V — конечное непустое множество *действий*. $\omega \in V^*$ — *трасса*. $L \subseteq V^*$ — неупорядоченное конечное множество трасс, *лог*.

Определение (Отношение порядка между действиями в логе)**.** Пусть L — лог над V. Пусть $A, B \in V$:

- $A >_L B$, если $\exists \omega = C_1 \dots C_n$, $C_i \in V$, $i \in \{1, \dots, n-1\}$ такие, что $\omega \in L$ и $C_1 = A$, $C_{i+1} = B$ — отношение предшествия, A *предшествует* B;
- $A \to_L B$, если $A >_L B$ и $B \not>_L A$ — прямое каузальное отношение;
- $A \#_L B$, если $A \not>_L B$ и $B \not>_L A$ — действия не встречаются вместе;
- $A ||_L B$, если $A >_L B$ и $B >_L A$ — потенциальный параллелизм.

Определение (Связь между переходами сети Петри и действиями лога)**.** Связь между переходами сети Петри $N = (P, T, F)$ и действиями лога $L \subseteq V^*$ задается сюръективным отображением $\tau : T \to V$.

Традиционно предполагается[8], что отображение τ является биективным, однако, в рамках настоящей работы рассматриваются лог и более широкий класс сетей Петри такие, что $|T| \geq |V|$.

Определение (Свойство полноты лога)**.** Пусть $N = (P, T, F)$ — надежная сеть Петри, L — лог над V. Отображение $\tau : T \to V$ задает связь между переходами сети Петри N и действиями лога L.

L — *лог сети* N, если для каждой трассы лога $\omega = \tau(t_1), \dots, \tau(t_n) \in L$, соответствующая последовательность переходов $\sigma = t_1, \dots, t_n$ — является срабатывающей последовательностью в N, начиная с разметки $[i]$ и заканчивая разметкой $[o]$. То есть $(N, [i])[\sigma\rangle(N, [o])$.

L — *полный лог сети* N, если:

1. L — лог рабочего процесса;
2. для любого другого лога рабочего процесса L' сети N выполнено: $>_{L'} \subseteq >_L$;
3. $\forall t \in T \; \exists \omega \in L$: $\tau(t) \in \omega$ — все переходы покрываются некоторой срабатывающей последовательностью.

Это условие корректности описывает связь между логом и сетью Петри, в частности, сможет ли построенная сеть Петри породить тот же лог, из которого данная сеть была построена. Если построенная модель описывает более широкий класс трасс, часть которых не описывает легальное поведение системы, то задача поиска аномалий не может быть корректно сформулирована, так как теряется возможность различить легальные трассы от аномальных.

Запрещенные конструкции для сетей Петри показаны на Рис. 4. Эти конструкции соответствуют ситуациям, когда происходят подряд выбор и синхронизация. Было показано[9], что наличие таких конструкций не только

[8] *Van der Aalst, W.* Workflow mining: Discovering process models from event logs / W. Van der Aalst, T. Weijters, L. Maruster // IEEE transactions on knowledge and data engineering. 2004. Т. 16, № 9. С. 1128—1142.

[9] *Mendling, J.* Seven process modeling guidelines (7PMG) / J. Mendling, H. A. Reijers, W. M. van der Aalst // Information and Software Technology. 2010. Т. 52, № 2. С. 127—136.

ухудшает восприятие человеком модели, но и усложняет процесс отладки модели. Такие модели имеют меньшую степень обоснованности и достоверности, а без этих предположений решать задачу поиска аномалий не представляется возможным. Так как если нет уверенности в том, что модель достаточно полно и при этом неизбыточно описывает все возможные легальные исполнения процесса, то и нет уверенности в том, что если произвольная трасса соответствует модели, то эта трасса точно не является аномалией. Под *выбором* в контексте сетей Петри понимается конструкция, когда из одного места существуют ребра в несколько переходов. Под *синхронизацией*, когда из нескольких мест есть ребра в один и тот же переход.

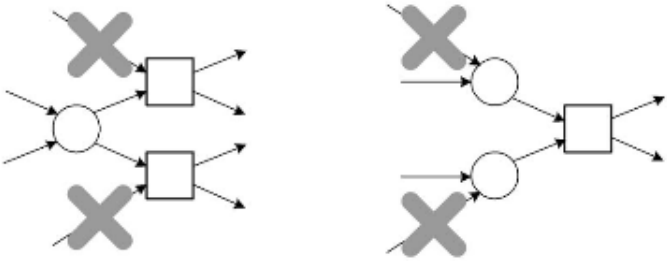

Рис. 4 — Запрещенные конструкции в сетях Петри

Определение (Неявное место)**.** Пусть $N = (P,T,F)$ — сеть Петри с начальной разметкой s. Место $p \in P$ называется *неявным*, если для всех достижимых разметок $s' \in [N,s\rangle$ и переходов $t \in p\bullet$ выполнено: если $s' \supseteq \bullet t \backslash \{p\}$, то $s' \supseteq \bullet t$.

То есть место p будет неявным, если для каждой достижимой из s разметки s' и переходов ω, которые следуют после этого места p, справедливо: если в s' есть метки в местах, предшествующих переходу ω, то в месте p в этой разметке тоже будет метка.

Следующее определение формализует интуитивные предположения выше о запрещенных конструкциях, тем самым определяя новый более узкий класс подходящих сетей Петри, которые являются подклассом сетей свободного выбора[10]:

Определение (Структурированная сеть Петри (SWF-сеть))**.** WF-сеть $N = (P,T,F)$ — SWF-сеть, если $\forall p \in P, \forall t \in T$, где $(p,t) \in F$ выполнено:

1. если $|p\bullet| > 1$, то $|\bullet t| = 1$ (левый рисунок на Рис. 4);
2. если $|\bullet t| > 1$, то $|\bullet p| = 1$ (правый рисунок на Рис. 4);

[10] *Desel, J.* Free choice Petri nets / J. Desel, J. Esparza. Cambridge university press, 1995.

3. в сети нет неявных мест.

Известно[11], что для SWF сетей существует алгоритм, определяющий, надежна ли сеть, работающий за полиномиальное время.

В этой главе проведен анализ требований, предъявляемых к сетям Петри, для решения задачи поиска аномалий, включая:

- свойство надежности;
- свойство, позволяющее обеспечить запрет на последовательное выполнение в сети конструкций синхронизации и выбора;
- полноту лога рабочего процесса для рассматриваемой сети;
- сохранение в сети отношения каузальности между переходами, если между соответствующими действиями в логе это отношение было выполнено.

Приведен пример простого процесса, состоящего из 4 действий, $L = \{ACD, AACD, ABCD\}$. В процессе есть повтор действия A и конструкция "ИЛИ", допускающая переход без выполнения какой-либо действия (трасса AC). Для этого процесса рассмотрены 5 моделей, $\mathbb{N}_1$, $\mathbb{N}_2$, $\mathbb{N}_3$, $\mathbb{N}_4$, $\mathbb{N}_5$, представленных в виде сети Петри. Доказан ряд утверждений относительно упомянутых выше свойств, которые задают базовые требования корректных сетей Петри для дальнейшего решения задачи поиска аномалий. Показано, почему каждую из моделей нельзя назвать подходящей под эти базовые требования. В таблице Теоремы 1 приведены результаты доказательств утверждений о свойствах корректности для рассмотренных моделей.

Теорема 1. *Результаты главы 2 могут быть представлены с помощью следующей таблицы:*

Сеть Петри	*Надежность*	*SWF-сеть*	*Полнота лога*	*Т. о кауз.*
$\mathbb{N}_1$	$+$	$-$	$-$	$+$
$\mathbb{N}_3$	$+$	$-$	$+$	$-$
$\mathbb{N}_4$	$-$	$+$	$-$	$-$
$\mathbb{N}_5$	$+$	$+$	$+$	$-$

Показано, что условия Теоремы о каузальности[12] не выполняются, когда в сети Петри есть переходы с одинаковыми действиями, однако, возможно добиться выполнения других свойств корректности с помощью перехода к более широкому классу сетей Петри (нетождественного определения отображения между переходами сети Петри и действиями лога τ). На примере сети $\mathbb{N}_2$ показано, что Теорема о каузальности не выполняется и при наличии блоков *OR-split*, *OR-join*, *AND-split*, *AND-join*. Это говорит

[11] *Van der Aalst, W. M.* The application of Petri nets to workflow management / W. M. Van der Aalst // Journal of circuits, systems, and computers. 1998. Т. 8, № 01. С. 21—66.

[12] *Van der Aalst, W.* Workflow mining: Discovering process models from event logs / W. Van der Aalst, T. Weijters, L. Maruster // IEEE transactions on knowledge and data engineering. 2004. Т. 16, № 9. С. 1128—1142.

о том, что решение задачи поиска аномалий с использованием модели в виде сети Петри, возможно только для простых процессов без ветвлений (“ИЛИ”) и требований одновременного выполнения нескольких условий (“И”).

Таблица в Теореме 1 с точки зрения задачи поиска аномалий свидетельствует о том, что для некоторого процесса (на примере простого процесса, состоящего из 4 возможных действий) наложение нескольких необходимых для дальнейшего поиска аномалий условий корректности приводит к задаче поиска подходящей модели. При этом решение этой задачи является переборным[13] и не гарантирует успешного результата. Тем самым строго сформулировать задачу поиска аномалий, используя для этого модель в терминах сети Петри, не всегда удается.

Для эффективного решения задачи поиска аномалий необходимым условием является наличие модели, которая бы описывала поведение процесса. Более того, необходимо, чтобы данная модель восстанавливалась по полному логу единственным образом и при этом не порождала бы новых трасс, которых не было в исходном логе. Однако приведенный выше пример процесса показывает, что сложности возникают не только с построением единственно возможной модели процесса, но и с произвольной моделью, которая бы удовлетворяла базовым свойствам корректности, которые обычно налагаются на модель в терминах сетей Петри. Таким образом, результатом главы 2 является демонстрация сложности однозначного восстановления модели в терминах сетей Петри с наложением нескольких условий корректности. Этот факт не позволяет в дальнейшем решить задачу обнаружения аномалий.

Третья глава посвящена восстановлению экземпляров информационных технологий, используя аппарат ориентированных ациклических графов. Построены обобщения методов восстановления моделей процессов в случаях реализации или одновременного функционирования нескольких информационных технологий. Основные результаты главы представлены в публикациях [3; 6].

Далее вводится ряд определений по аналогии с работой[14]. Некоторые из определений были модифицированы для удобства их использования в задаче поиска аномалий.

Пусть задан конечный непустой алфавит V возможных действий для процесса P. Аналогично главе 2, определяются понятия трассы, как некоторого слова из V^*, понятия лога, как неупорядоченного конечного множества трасс.

[13] *Питерсон, Д.* Теория сетей Петри и моделирование систем / Д. Питерсон. Мир, 1984.

[14] *Agrawal, R.* Mining process models from workflow logs / R. Agrawal, D. Gunopulos, F. Leymann // International Conference on Extending Database Technology. Springer. 1998. С. 467—483.

Пример: Запись лога $L = \{ABC, DF\}$ означает, что в нем содержатся 2 трассы и множество действий $\{A, B, C, D, F\}$.

Вводится определение зависимости для лога. В модели процесса каждая зависимость представляется либо как направленная дуга, либо как направленный путь от одного действия к другому.

Определение (Частичный предпорядок для лога)**.** Пусть задан лог L одного процесса P. Действие B *следует* за действием A ($A \lesssim_L B$), если в логе L действие B начинается после того, как заканчивается выполнение действия A, либо существует действие C такое, что C следует за A, а B следует за C.

Например, для лога $L = \{ABA, DF\}$ справедливо: $A \lesssim_L B$, $B \lesssim_L A$, $A \lesssim_L A$, $D \lesssim_L F$.

Определение (Зависимость действий)**.** Пусть задан лог L одного процесса P. Действие B *зависит* от действия A ($A \rightarrow_L B$), если B следует за A, а A не следует за B. Если A следует за B и B следует за A, либо A не следует за B и B не следует за A, то A и B *независимые* действия.

Так, для лога $L = \{ABA, DF\}$ справедливо: $D \rightarrow_L F$.

Определения частичного предпорядка и зависимости действий могут быть заданы и для одной трассы. Далее, если не сказано обратного, записи $A \lesssim B$, $A \rightarrow B$ обозначают частичный предпорядок и зависимость действий в логе L.

Определение (Граф зависимостей)**.** Пусть задано множество действий V и лог L некоторого процесса, ориентированный граф G называется *графом зависимостей*, если ориентированный путь из A в B в графе G существует тогда и только тогда, когда B зависит от A.

Можно ввести определение *подграфа G', порождаемого трассой ω*, в графе зависимостей $G = (V,E) : G' = (V',E')$, где $V' = \{A \in V | A \in \omega\}$ и $E' = \{(B, A) \in E |\ B \lesssim_\omega A\}$, где $\lesssim_\omega$ — отношение частичного предпорядка на действиях, задаваемое трассой ω.

Определение (Согласованность трассы с графом зависимостей)**.** Пусть задан граф зависимостей процесса P $G = (V,E)$, трасса ω. Трасса ω согласуется с G, если подграф $G' = (V',E')$, порождаемый трассой ω, связен; первое и последнее действия трассы ω — действия, начинающее и завершающее процесс P; все вершины в V' достижимы из начального действия; никакая зависимость в графе G не нарушена упорядочиванием действий в ω.

Не все графы зависимостей, содержащие в себе зависимости некоторого лога L, являются корректными. Для формализации ограничений,

которые обычно налагаются на графы зависимостей, чтобы графы зависимостей могли называться моделями процесса, вводится определение *графа, конформного логу*:

Определение (Конформный логу граф)**.** Граф зависимостей G называется *конформным* логу L, если выполнены следующие условия:

- для каждой зависимости в L существует путь в G;
- между независимыми действиями в L не существует пути в G;
- каждая трасса лога L согласована с G.

Пусть задана трасса $\omega = A_1 \dots A_q$. *Подстрокой трассы* ω *длины* $r \leq q$ называется трасса $\omega' = A_i A_{i+1} \dots A_{i+r-1}$, где $i = 1, \dots, q - r + 1$.

В работе[15] были предложены три алгоритма и рассчитана временная сложность построения модели в виде ациклического ориентированного графа, в зависимости от различных свойств лога исполнений одного процесса. Пусть $m = |L|$ — количество трасс в логе, $n = |V|$ — количество возможных действий процесса P. Предполагается, что $m \to \infty$, $n \to \infty$. Основным предположением о входных параметрах алгоритмов является предположение, что количество трасс m в логе L превышает количество действий n, $m >> n$. Под элементарными операциями для вычисления оценок сложности алгоритмов понимаются простейшие операции редактирования над графами (например: добавление/удаление вершины, добавление/удаление ребра, слияние/расщепление вершин). Данные результаты были переформулированы для удобства их использования в задаче поиска аномалий и используются в настоящей главе в виде лемм:

Лемма 1 (О сложности построения конформного DAG специального вида)**.** *Пусть задан лог процесса L. Известно, что L содержит только трассы, в которых каждое действие алфавита V встречается ровно по одному разу. Известно, что конформный логу L граф G не содержит циклов. Сложность алгоритма построения конформного графа G составляет $O(mn^2)$.*

Лемма 2 (О сложности построения конформного DAG)**.** *Пусть задан лог процесса L. Известно, что конформный логу L граф G не содержит циклов. Сложность алгоритма построения конформного графа G составляет $O\left(mn^3\right)$.*

Лемма 3 (О сложности построения конформного графа)**.** *Пусть задан лог процесса L. Сложность алгоритма построения конформного логу L графа G составляет $O\left(m\left(kn\right)^3\right)$, где k – максимальное количество повторов некоторого действия в логе.*

[15] *Agrawal, R.* Mining process models from workflow logs / R. Agrawal, D. Gunopulos, F. Leymann // International Conference on Extending Database Technology. Springer. 1998. С. 467—483.

Основными этапами предложенных алгоритмов, которые строят модель процесса в виде графа по заданному логу L, являются:

1. добавление в граф всех дуг, соответствующих зависимостям между действиями, представленными в логе L;
2. удаление дуг внутри сильно связных компонент графа;
3. добавление меток на дуги, которые присутствуют в транзитивных замыканиях подграфов, порождаемых трассами лога L. Удаление дуг без меток;
4. склейка вершин, соответствующих одинаковым действиям.

Первая задача, результаты решения которой представлены в этой главе, является построение формальных моделей нескольких процессов, по заданному логу возможных исполнений этих процессов. Сначала рассматривается случай, когда известно, что одна трасса лога может содержать данные ровно одного процесса. Далее рассмотрен случай, когда в рамках одной трассы лога могут встречаться данные нескольких процессов.

Вторая задача, решение которой описано в этой главе, направлена на поиск аномалий в трассе с использованием построенных ранее моделей процессов. Рассматривается случай, когда проверяемая на аномальность трасса является исполнением только одного процесса, так и когда проверяемая на аномальность трасса содержит в себе исполнение нескольких процессов.

Рассматривается s процессов $P_1,\ldots,P_s$, с непустыми множествами действий $V_1,\ldots,V_s$ и графами $G_1 = (V_1,E_1),\ldots,G_s = (V_s,E_s)$ соответственно.

Пусть задан некоторый лог L, $M = |L| \to \infty$, $n = \max_i |V_i| \to \infty$, $i = 1,\ldots,s$, k – максимальное количество повторов действия в логе L. Также, как и в работе[16], предполагается, что $M >> n$, количество процессов s процессов конечно.

Теорема 2 (Процессы с различными множествами действий)**.** *Пусть задан лог L для s процессов. Пусть выполнены условия Лемм 1, 2, 3. Пусть s процессов имеют попарно различные множества действий, $V_i \cap V_j = \emptyset$, при $i \neq j, i, j = 1,\ldots,s$. Тогда сложность построения s конформных графов составляет $O\left(Mn^2\right), O\left(Mn^3\right), O\left(M\left(kn\right)^3\right)$ соответственно, в зависимости от свойств лога L.*

Если построено s конформных графов зависимостей для s процессов и известно, что в поступающей трассе ω может быть исполнение только одного процесса, то задача поиска аномалий в ω сводится к задаче определения согласованности этой трассы с s графами зависимостей.

Пусть $n = \max_i |V_i|$, $e = \max_i |E_i|, i = 1,\ldots,s$, $u = |\omega|$ — длина трассы ω, $u \to \infty$.

[16] *Agrawal, R.* Mining process models from workflow logs / R. Agrawal, D. Gunopulos, F. Leymann // International Conference on Extending Database Technology. Springer. 1998. С. 467—483.

Теорема 3 (Поиск аномалий в процессах с различными множествами действий)**.** *Пусть s процессов имеют попарно различные множества действий, $V_i \cap V_j = \emptyset$, при $i \neq j, i,\ j = 1, \ldots, s$. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за не более, чем $s + C_1 un + C_2 en^2$ операций, где C_1, C_2 — некоторые константы.*

Если для задачи поиска аномалий, построены модели процессов в виде ациклических ориентированных графов, тогда параметры s, n, e будут константами и справедливо:

Следствие 1 (Поиск аномалий в процессах с различными множествами действий)**.** *Пусть s процессов имеют попарно различные множества действий, $V_i \cap V_j = \emptyset$, при $i \neq j, i, j = 1, \ldots, s$. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за $O(u)$.*

Пусть $n = \max_i |V_i| \to \infty$, $i = 1, \ldots, s$. Пусть $m = \max_i |L_i| \to \infty$, L_i – часть лога L, соответствующая i-му процессу, $i = 1, \ldots, s$, k — максимальное количество повторов действия в логе L. Если известно, что процесс однозначно можно идентифицировать его начальным действием, то при условии s различных начальных действий, справедлива следующая Теорема:

Теорема 4 (Отличимые процессы по начальному действию)**.** *Пусть задан лог L для s процессов. Пусть s процессов имеют попарно различные начальные действия. Тогда сложность построения s конформных графов составляет $O\left(mn^2\right)$, $O\left(mn^3\right)$, $O\left(m\left(kn\right)^3\right)$ соответственно, в зависимости от свойств лога L.*

Пусть $n = \max_i |V_i|$, $e = \max_i |E_i|$,$i = 1, \ldots, s$, $u = |\omega|$ — длина трассы ω, $u \to \infty$.

Теорема 5 (Поиск аномалий в процессах, отличимых по начальному действию)**.** *Пусть s процессов имеют попарно различные начальные действия. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за не более, чем $s + C_1 un + C_2 en^2$ операций, где C_1, C_2 — некоторые константы.*

Если для задачи поиска аномалий, построены модели процессов в виде ациклических ориентированных графов, тогда параметры s, n, e будут константами и справедливо:

Следствие 2 (Поиск аномалий в процессах, отличимых по начальному действию)**.** *Пусть s процессов имеют попарно различные начальные действия. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за $O(u)$.*

Рассмотрен случай, когда не налагается условие на попарно различные начальные действия для каждого из s процессов, но есть возможность разделить лог L на части по одному уникальному для процесса действию. Таким образом, обеспечивается необходимое и достаточное условие существования системы различных представителей для теоремы Ф. Холла. Пусть $n = \max_i |V_i| \to \infty$, $i = 1, \ldots, s$. Пусть $m = \max_i |L_i| \to \infty$, L_i – часть лога L, соответствующая i-му процессу, $i = 1, \ldots, s$, k — максимальное количество повторов действия в логе L.

Теорема 6 (Отличимые процессы по некоторому действию)**.** *Пусть задан лог L для s процессов. Пусть каждый из s процессов имеет по одному известному действию A_i, $i = 1, \ldots, s$ такому, что оно не содержится в остальных процессах и содержится в каждой трассе, соответствующей i-му процессу. Тогда сложность построения s конформных графов составляет $O\left(mn^2\right)$, $O\left(mn^3\right)$, $O\left(m\left(kn\right)^3\right)$ соответственно, в зависимости от свойств лога L.*

Пусть $n = \max_i |V_i|$, $e = \max_i |E_i|$,$i = 1, \ldots,s$, $u = |\omega| \to \infty$ — длина трассы ω.

Теорема 7 (Поиск аномалий в процессах, отличимых по некоторому действию)**.** *Пусть каждый из s процессов имеет по одному известному действию A_i, $i = 1, \ldots, s$ такому, что оно не содержится в остальных процессах и содержится в каждой трассе, соответствующей i-му процессу. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за не более, чем $s + C_1 un + C_2 en^2$ операций, где C_1, C_2 — некоторые константы.*

Если для задачи поиска аномалий, построены модели процессов в виде ациклических ориентированных графов, тогда параметры s, n, e будут константами и справедливо:

Следствие 3 (Поиск аномалий в процессах, отличимых по некоторому действию)**.** *Пусть каждый из s процессов имеет по одному известному действию A_i, $i = 1, \ldots, s$ такому, что оно не содержится в остальных процессах и содержится в каждой трассе, соответствующей i-му процессу. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в некоторой трассе ω может быть выполнен за $O\left(u\right)$.*

Теорема 4 рассматривает случай уникальных для процессов подстрок длины 1, которые начинаются с первой буквы трасс лога L. Пусть $m = \max_i |L_i| \to \infty, n = \max_i |V_i| \to \infty$, $i = 1, \ldots, s$, k — максимальное количество повторов действия в логе L. Справедливо обобщение этой теоремы:

Теорема 8 (Отличимые процессы по конечной подстроке)**.** *Пусть задан лог L для s процессов. Пусть известно, что каждая из трасс лога, начиная с некоторого номера i содержит подстроки длины r, каждая из которых может принадлежать только одному из s процессов. Тогда сложность построения s конформных графов составляет $O\left(mn^2\right)$, $O\left(mn^3\right)$, $O\left(m\left(kn\right)^3\right)$ соответственно, в зависимости от свойств лога L.*

Пусть $n = \max_i |V_i|$, $e = \max_i |E_i|$,$i = 1, \ldots, s$, $u = |\omega| \to \infty$ — длина трассы ω.

Теорема 9 (Поиск аномалий в процессах, отличимых по конечной подстроке)**.** *Пусть известно, что трасса ω, начиная с некоторого номера i содержит подстроку длины r, которая может принадлежать только одному из s процессов. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в трассе ω может быть выполнен за не более, чем $C_1 un + C_2 en^2$ операций, где C_1, C_2 — некоторые константы.*

Если для задачи поиска аномалий, построены модели процессов в виде ациклических ориентированных графов, тогда параметры n, e будут константами и справедливо:

Следствие 4 (Поиск аномалий в процессах, отличимых по конечной подстроке)**.** *Пусть известно, что трасса ω, начиная с некоторого номера i содержит подстроку длины r, которая может принадлежать только одному из s процессов. Пусть для s процессов построено s конформных графов зависимостей, тогда поиск аномалий в трассе ω может быть выполнен за $O\left(u\right)$.*

Пусть теперь в логе L в одной трассе могут содержаться данные нескольких процессов.

Пусть $M = |L| \to \infty$, $n = \max_i |V_i| \to \infty$, $i = 1, \ldots, s$, $M >> n$. Пусть множества действий для процессов не пересекаются, то есть $V_i \cap V_j = \varnothing$, при $i \neq j$, $i, j = 1, \ldots, s$.

Теорема 10 (Множество процессов в трассе без повторов действий)**.** *Пусть задан лог L для s процессов. В каждой из трасс лога L могут присутствовать действия от 1 до s процессов и в рамках одной трассы нет повторов действий. Пусть конформные графы для s процессов не содержат циклов. Пусть множества действий для процессов не пересекаются, то есть $V_i \cap V_j = \varnothing$, при $i \neq j$, $i, j = 1, \ldots, s$. Если для всех действий $A \in V_i, B \in V_j, i \neq j$, i, $j = 1, \ldots,$ s в логе L выполнено: $A \lesssim B$ и $B \lesssim A$, то сложность построения s конформных графов составляет $O\left(Mn^2\right)$.*

Если построенные s конформных графов по Теореме 10 обладают свойствами из Теорем 2, 4, 6, 8 и некоторая трасса ω содержит в себе исполнение только одного процесса, то для ответа на вопрос, является ли трасса ω аномальной, могут быть применимы Следствия 1, 2, 3, 4 соответственно.

Следующая теорема отвечает на вопрос о сложности поиска аномалий, когда и поступающая трасса ω содержит в себе исполнения нескольких процессов:

Теорема 11 (Поиск аномалий для трассы, содержащей исполнения множества процессов, имеющих различные множества действий)**.** *Пусть множества действий для s процессов не пересекаются, то есть $V_i \cap V_j = \varnothing$, при $i \neq j$, $i, j = 1, \ldots, s$. Пусть для s процессов построено s конформных графов зависимостей. Пусть в некоторой трассе ω могут содержаться исполнения множества процессов. Тогда поиск аномалий в трассе ω может быть выполнен за не более, чем $s(s + C_1 un + C_2 en^2)$ операций, где C_1, C_2 – некоторые константы.*

Если для задачи поиска аномалий построены модели процессов в виде ациклических ориентированных графов, тогда параметры s, n, e будут константами и справедливо:

Следствие 5 (Поиск аномалий для трассы, содержащей исполнения множества процессов, имеющих различные множества действий)**.** *Пусть множества действий для s процессов не пересекаются, то есть $V_i \cap V_j = \varnothing$, при $i \neq j$, $i, j = 1, \ldots, s$. Пусть для s процессов построено s конформных графов зависимостей. Пусть в некоторой трассе ω могут содержаться исполнения множества процессов. Тогда поиск аномалий в трассе ω может быть выполнен за $O(u)$.*

Можно рассчитать нижнюю границу количества действий в логе L по всем трассам для ограничений, налагаемых в Теореме 10. Пусть $R = \sum_{\omega \in L} |\omega|$ — общее количество действий в логе, $n_i = |V_i|, i = 1, \ldots, s$ — количество действий каждого из s процессов.

Следствие 6. *Пусть задан лог L, который не содержит повторов действий. Пусть в каждой из трасс лога L могут присутствовать действия от 1 до s процессов. Пусть для всех действий $A \in V_i, B \in V_j, i \neq j, i, j = 1, \ldots, s$ в логе L выполнено: $A \lesssim B$ и $B \lesssim A$. Тогда $|L| \geq 2$, $R \geq 2 \sum_{i=1}^{s} n_i$.*

В этой главе рассмотрена задача построения формальной модели множества процессов в виде ациклического ориентированного графа и задача поиска аномалий с помощью построенных эталонных моделей процесса. Найдены ограничения на возможности применения алгоритмов для одного процесса и определены оценки сложности построения формальных моделей для множества процессов в зависимости от свойств лога.

Для задачи построения моделей процесса рассмотрены случаи наличия исполнений нескольких процессов как между трассами одного лога, Теоремы 2, 4, 6, 8, так и внутри трасс одного лога, Теорема 10.

Показано, что с помощью системы различных представителей возможно эффективно выделять траектории множества различных процессов, Теорема 6.

Для задачи поиска аномалий предложены алгоритмы ее решения. Показано, что оценка сложности предложенных алгоритмов является линейной относительно длины трассы, для которой требуется дать ответ об аномальности, и квадратичной относительно максимально возможного количества действий и зависимостей в одном моделируемом процессе, Теоремы 3, 5, 7, 9, 11.

В **Заключении** приведены основные результаты работы, которые заключаются в следующем.

1. Исследована возможность использования математических моделей в виде ациклических ориентированных графов (DAG) для решения задачи построения модели процесса и для решения задачи поиска аномалий.

   Показано, что для моделей, сформулированных в терминах ациклических ориентированных графов, удается успешно решать задачу поиска аномалий при условии наличия нескольких одновременно функционирующих процессов. В том числе продемонстрировано, что с помощью системы различных представителей можно эффективно выделять траектории множества различных процессов.
2. Получены временные оценки сложности построения моделей процесса при описании модели процесса с помощью ациклических ориентированных графов.

   При этом продемонстрировано, что в ряде случаев алгоритмы восстановления нескольких одновременно функционирующих процессов полностью повторяют их аналоги для одного функционирующего процесса, что позволяет использовать уже готовые оценки сложности для решения задачи поиска аномалий.
3. Получены временные оценки сложности выявления аномалий при описании модели процесса с помощью ациклических ориентированных графов, которые являются линейными относительно длины проверяемой трассы.
4. Показано, что использование формального аппарата в виде сетей Петри для решения задачи поиска аномалий затруднено тем фактом, что модель процесса не всегда удается однозначно восстановить.

   Налагая простые требования корректности, такие, как: свойство надежности; свойство, позволяющее обеспечить запрет на последовательное выполнение в сети конструкций синхронизации и выбора;

полнота лога рабочего процесса для рассматриваемой сети; сохранение в сети отношения каузальности между переходами, если между соответствующими действиями в логе это отношение было выполнено — не гарантирован результат построения хоть какой-либо подходящей однозначной, корректной модели процесса.

Продемонстрировано, что без перехода к более сложному классу сетей Петри класс технологических процессов, для которого возможно решить задачу поиска аномалий с использованием восстановленной моделью процесса, сильно ограничен.

5. Показано, что не каждая математическая модель описания реального процесса является подходящей для эффективного решения задач информационной безопасности. Тем самым продемонстрирован подход, позволяющий выбирать модели для конкретных задач информационной безопасности.

Благодарности

Автор выражает благодарность и большую признательность своему научному руководителю доктору физико-математических наук, профессору Грушо Александру Александровичу за постановку задач, а также сотрудникам кафедры информационной безопасности факультета вычислительной математики и кибернетики МГУ имени М. В. Ломоносова за внимание к работе.

Автор благодарит семью, друзей и коллег за оказанную поддержку в процессе написания работы.

Публикации автора по теме диссертации

В рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 2.3.6

1. Построение моделей процесса с помощью простых сетей Петри / И. Ю. Терёхина, А. А. Грушо, Е. Е. Тимонина, С. Я. Шоргин // Системы и средства информатики. — 2020. — Т. 30, № 4. — С. 61—75. — (ВАК, RSCI, Двухлетний импакт-фактор РИНЦ 2022 – 0.556) // Соавторам принадлежат постановка задачи и проверка результатов. Остальные результаты статьи получены Терёхиной И.Ю.
2. Выявление аномалий с помощью метаданных / А. А. Грушо, Е. Е. Тимонина, Н. А. Грушо, И. Ю. Терёхина // Информатика и ее применения. — 2020. — Т. 14, № 3. — С. 76—80. — (Scopus, ВАК, RSCI, SJR – 0.22, Двухлетний импакт-фактор РИНЦ 2022 – 0.787) // Соавторам принадлежат постановка задачи, проверка результатов статьи, а также результаты по Главам 1, 3, 4. Вклад Терёхиной И.Ю. состоит в Главе 2, Теореме 2 и ее следствии – о необходимом и достаточном условии наличия элементов СРП в функционировании различных ИТ.
3. *Терёхина, И. Ю.* Выявление аномалий с использованием построенной модели процессов в виде ациклического ориентированного графа / И. Ю. Терёхина // Программная инженерия. — 2023. — Т. 14, № 6. — С. 285—291. — (ВАК, RSCI, Двухлетний импакт-фактор РИНЦ 2022 – 0,539).

В сборниках трудов конференций

4. *Терёхина, И. Ю.* О поиске аномалий с использованием построенных моделей нескольких процессов / И. Ю. Терёхина // Ломоносовские чтения-2023: научная конференция, факультет ВМК МГУ имени М.В.Ломоносова. Тезисы докладов. — Москва : ООО МАКС Пресс, 2023. — С. 183—184.
5. *Terekhina, I. Y.* Выявление аномалий в условиях функционирующих процессов / I. Y. Terekhina // Proceedings of Academician O.B. Lupanov 14th International Scientific Seminar "Discrete Mathematics and Its Applications". — Keldysh Institute of Applied Mathematics, 2022. — С. 288—291.

В прочих изданиях

6. *Teryokhina, I.* Anomaly detection in several running processes / I. Teryokhina // International Journal of Open Information Technologies. — 2022. — Т. 10, № 1. — С. 21—27. — (ВАК, Двухлетний импакт-фактор РИНЦ 2022 – 0.802).

Терёхина Ирина Юрьевна

Методы выявления аномалий в условиях смеси технологических процессов, сопровождающих наблюдаемый объект

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать 23.09.2024. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Отдел полиграфии Научной библиотеки МГУ имени М.В. Ломоносова
119192, Москва, Ломоносовский проспект, 27.