DOI:

ПОДГОТОВКА И ОБРАБОТКА БОЛЬШИХ ДАННЫХ В СИСТЕМАХ АДАПТИВНОГО МОНИТОРИНГА БЕЗОПАСНОСТИ КФС

Полтавцева М.А.

Санкт – Петербургский политехнический университет Петра Великого,
Россия, г. Санкт-Петербург ул. Политехническая д.29
poltavtseva@ibks.spbstu.ru

Аннотация: Рассматривается задача адаптивного мониторинга безопасности киберфизических систем. В работе приводится архитектура системы, порядок сбора и подготовки данных, алгоритм иерархической агрегации. Дан авторский подход к многомерному рассмотрению объекта защиты в условиях множества разнородных методов анализа и результаты оценки эффективности.

Ключевые слова: Большие данные, информационная безопасность, мониторинг безопасности, адаптивный мониторинг, киберфизические системы, многомерная агрегация данных, многомерный анализ.

Введение

В связи с развитием систем цифрового производства, принятой стратегией научно-технологического развития Российской Федерации и программой «Цифровая экономика Российской Федерации» сегодня происходит переход к цифровым, интеллектуальным производственным технологиям. Это процесс связан с широким внедрением в промышленность киберфизических объектов и построение на их основе киберфизических систем (КФС), в том числе, распределенного характера.

Киберфизические объекты отличаются сочетанием цифрового управления и коммуникаций с контролем физического процесса. Киберфизические системы уже применяются в энергетике, медицине, системах водоснабжения и многих других областях. Однако, сегодня такие системы также широко подвержены атакам злоумышленников. Только за 2019 год можно отметить атаки на энергетические объекты Венесуэлы через автоматическую систему контроля ГЭС, повлекшие массовое отключение электроснабжения по всей стране. Схожий инцидент произошел в ЮАР. Из-за кибератаки подразделения Rheinmetall приостановили работу в трех странах. Согласно отчету Positive Technologies в 2019 году для атак на промышленные компании выросла с 4% до 10% в общем числе атак, а общее число превысило показатели 2018 года на 20%. При этом более 90% таких атак проводятся с помощью компьютерного оборудования, сетевых коммуникаций [1].

Сложившаяся ситуация повышает актуальность развития и применения систем управления информационной безопасностью в области распределенных киберфизических решений. Эффективность управления безопасностью зависит от полноты, достоверности и своевременности информирования. Эта задача возложена на системы мониторинга безопасности. Данная работа посвящена задаче построения подсистемы сбора, подготовки и обработки данных для обеспечения полноты и своевременности мониторинга безопасности распределенных киберфизических систем.

1 Мониторинг безопасности распределенных киберфизических систем

1.1 Проблемы мониторинга безопасности распределенных киберфизических систем

Сегодня, в эпоху Больших данных и больших цифровых систем, мониторинг безопасности распределенных киберфизических объектов стакивается с целым рядом проблем, вызванных новыми типами атак, процессами цифровизации производства и роста скорости и объема поступления данных в систему мониторинга.

Основным типом нападений на промышленные системы являются целенаправленные (APT) атаки. Этот типа атак всегда был лидером в области нападений на промышленные предприятия и объекты. В 2019 году для таких атак не только в этой области, но и в целом среди всех зарегистрированных инцидентов составила 60% [1]. При этом фигурируют различные цели атаки:

- влияние на промышленные и производственные процессы;
- кража информации (персональные и учетные данные, данные о физических процессах и т.д.);
- перехват управления отдельных устройств и мобильных объектов (например, дронов);
- заражение вредоносным ПО и вымогательство;

И другие. Целенаправленные атаки характеризуются уникальностью, разнообразием методов, часто-длительностью во времени, длинными цепочками действий злоумышленника, использованием как технических подходов, так и социальной инженерии. Разнообразие целей и методов воздействий

приводит к необходимости рассмотрения с точки зрения безопасности не только задачи обнаружения вторжений или атак на отказ в обслуживании (DDoS). Система управления информационной безопасностью должна иметь возможность оценки исторических данных, расследования инцидентов, сопоставления инцидентов в разных частях и на разных уровнях объекта защиты, оценивать информированность злоумышленника.

Таким образом, современная ситуация в области мониторинга безопасности распределенных киберфизических систем характеризуется рядом проблем:

1. Разнообразием целей и задач мониторинга;
2. Семантической и структурной гетерогенностью анализируемых данных (даже в рамках одной задачи);
3. Ростом объема и скорости поступления данных.

Сложность и комплексность атак на промышленные киберфизические системы обуславливает как разнообразие целей мониторинга, о чем было сказано выше, так и необходимость совместного анализа семантически гетерогенных данных, в частности, коммуникационных данных сетевого трафика и конечных данных – показателей физических процессов, информации датчиков и сенсоров киберфизических объектов.

В свою очередь путь максимальной детализации и гранулирования данных приводит к их лавинообразному росту, а значит большим вычислительным и, главное, временным затратам на обработку. С одной стороны, это ведет к нарушению своевременности информирования, а с другой – зачастую не приводит к увеличению точности обнаружения инцидентов, ложным срабатываниям и т.д.

Такая ситуация приводит к необходимости создания методов и средств подготовки и обработки данных в системе мониторинга безопасности позволяющих решить проблему гетерогенности и роста данных за счет их предварительной обработки.

1.2 Особенности адаптивного мониторинга распределенных киберфизических систем

Разработка эффективных методов мониторинга распределенных киберфизических систем (КФС) невозможна без учета их специфики как объекта защиты. Совмещение цифровых технологий, современной коммуникационной среды и управления физическими процессами обуславливает целый ряд особенностей, которые, с одной стороны, должны быть учтены при сборе, подготовке и обработке данных, а с другой – позволят организовать эти процессы более эффективно. Ключевые особенности КФС и данных их мониторинга безопасности можно определить, как:

1. Сочетание коммуникационных и конечных данных при обнаружении вторжений и поиске аномалий работы;
2. Динамическое изменение методов анализа;
3. Динамическое изменение среды и компонентов;
4. Периодичность и само подобие протекающих процессов (как в коммуникационной среде, так и в физических процессах) [2].

Рассмотрим эти аспекты более подробно.

Работы в области мониторинга современных промышленных КФС отличаются вопросом оценки конечных данных физических устройств [3] или сетевого трафика [4]. Это приводит к тенденции по совместной оценке конечных данных, выявлению корреляционных и иных зависимостей [5]. Следовательно, данные обоих видов должны при обработке представляться единообразно и должна существовать возможность их совместной обработки и оценки.

Развитие цифровых производственных систем приводит как к увеличению числа и разнообразия атак на них, так и к совершенствованию методов защиты, на фоне динамического изменения технологий и программно-аппаратной базы. Этот факт выдвигает требование адаптивности к системам мониторинга безопасности, причем адаптивность в данном случае понимается и как адаптивность к изменениям внешней среды, так и к внутренним компонентам системы. Адаптивная система мониторинга информационной безопасности должна обеспечивать:

1. Гибкость системы мониторинга к изменениям в компонентах системы и процессах их взаимодействия.
2. Адаптивность к новым атакам и, как следствие, новым методам защиты и обнаружения.

Эти требования обуславливают адаптивность системы сбора и подготовки данных. В результате ее работы данные должные представляться в виде, пригодном для применения основных методов и подходов современного анализа, возможности определения взаимосвязей и корректного совместного рассмотрения различных семантически гетерогенных параметров (даже если это рассмотрение не требуется в данный момент). Также необходимо учитывать возможность рассмотрения одних и тех же

(или различных) наборов параметров на разных временных промежутках, для применения оценок на основе самоподобия и фракталов.

Приведенные аспекты формируют специфические требования к подсистеме сбора, подготовки и обработки данных в системе мониторинга безопасности КФС. Это требование унификации представления само подобных гетерогенных данных, использование модели данных на адаптированной для совместного анализа и рассмотрения данных в самых разных ракурсах и на различных временных промежутках, требование поддержания нескольких наборов данных каждого параметра в соответствии с различными периодами времени. К ним добавляются общие требования к системе мониторинга, такие как заданное время реакции, возможность обработки потоковых и исторических данных.

2 Подготовка и обработка данных в системе адаптивного мониторинга безопасности

2.1. Архитектура системы адаптивного мониторинга

Приведенные требования обуславливают общую архитектуру системы мониторинга безопасности распределенных киберфизических систем, включающую компоненты сбора данных, подготовки и обработки (анализа) информации. Ключевыми аспектами являются компоненты интеллектуального управления.

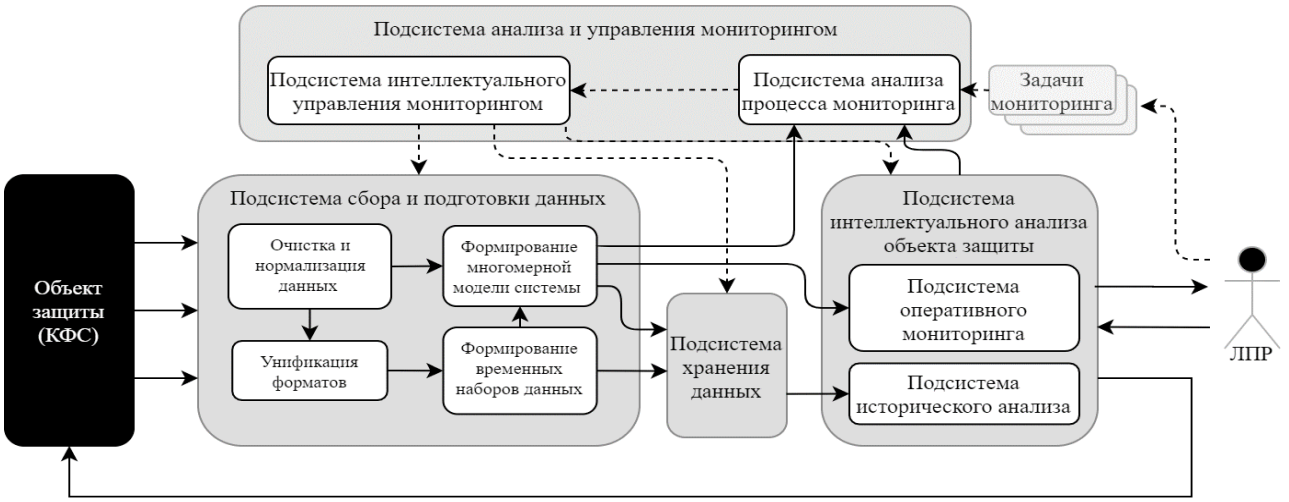

Рис.1. Архитектура системы адаптивного мониторинга

Особенностями предложенной архитектуры являются наличие подсистемы анализа и управления сами процессом мониторинга, а также некоторые компоненты системы сбора и подготовки данных. Подсистема анализа и управления мониторингом включает компоненты анализа и управления, основой для функционирования которых являются данные подсистемы интеллектуального анализа. Стоит отметить, что в этот набор данных входят как результаты применения аналитических функций, так и необходимые параметры многомерной модели системы.

Подсистема сбора и подготовки данных для мониторинга безопасности распределенных киберфизических систем включает дополнительные модули. Помимо модуля очистки и нормализации данных[6] вводятся модули унификации форматов, формирования временных наборов данных и многомерной модели системы.

2.2 Сбор и подготовка данных

Сбор и подготовка данных для распределенных киберфизических систем включает сбор и оценку двух связанных потоков данных:

1. Данные конечных устройств, характеризующие протекание физических процессов в объекте защиты;
2. Данные коммуникационные, характеризующие цифровые коммуникации объекта.

Данные конечных устройств включают в себя физические показатели датчиков и сенсоров, а также – управляющие команды и воздействия, поступающие на системы управления технологическими процессами. Коммуникационные данные представляют собой дампы трафика, пересылаемые маршрутизаторами напрямую или после локального разбора и анализа. Общей особенностью этих видов данных является их дальнейшее представление в виде временных рядов множества параметров.

На начальном этапе все данные представляются в виде кортежей $< Timestamp, \{Parameter\} >$, состоящих из временной метки и набора связанных с ней параметров. При этом каждое событие в системе (генерация данных, приход пакета и т.д.) порождает некоторое количество таких кортежей.

После процесса нормализации, более подробно представленного в [6], данные приводятся у унифицированному формату на основе приведенного кортежа и формируются временные ряды на основе меток *Timestamp*.

Основной задачей в этом случае становится организация процесса подготовки данных и формирования временных рядов в условиях интенсивного внешнего потока поступающей информации. Для этого используются основные технологии повышения производительности вычислений и операций над данными, такие как:

1. Конвейеризация задач обработки данных;
2. Построение параллельной системы обработки данных [7].

Технологии вычислительного конвейера позволяют повысить скорость обработки информации. Для построения параллельной обработки используются алгоритмы управления нагрузкой, минимизирующие пересылку данных между узлами – обработчиками. Такой подход позволяет повысить производительность системы подготовки данных и сократить время обработки сырой информации.

Для того, чтобы повысить эффективность последующих операций анализа данных необходимо организовать агрегацию данных на этапе подготовки с учетом следующих требований:

1. Каждый вид параметров (или набор из нескольких параметров) может агрегироваться на основании нескольких временных промежутков с разным шагом агрегации;
2. Пересылка данных между временными рядами и узлом-генератором данных для агрегации должна быть минимизирована;
3. Размер (с точки зрения объема данных) итоговых временных рядов должен быть минимизирован для ускорения анализа.

Для выполнения этих требований применяется метод иерархической агрегации данных, основанных на введении над временными рядами отношения наследования в виде PCR (Public-Child-Relation) связи [8]. Алгоритм реализации иерархической агрегации временных рядов в системе адаптивного мониторинга приведен на рисунке 2.

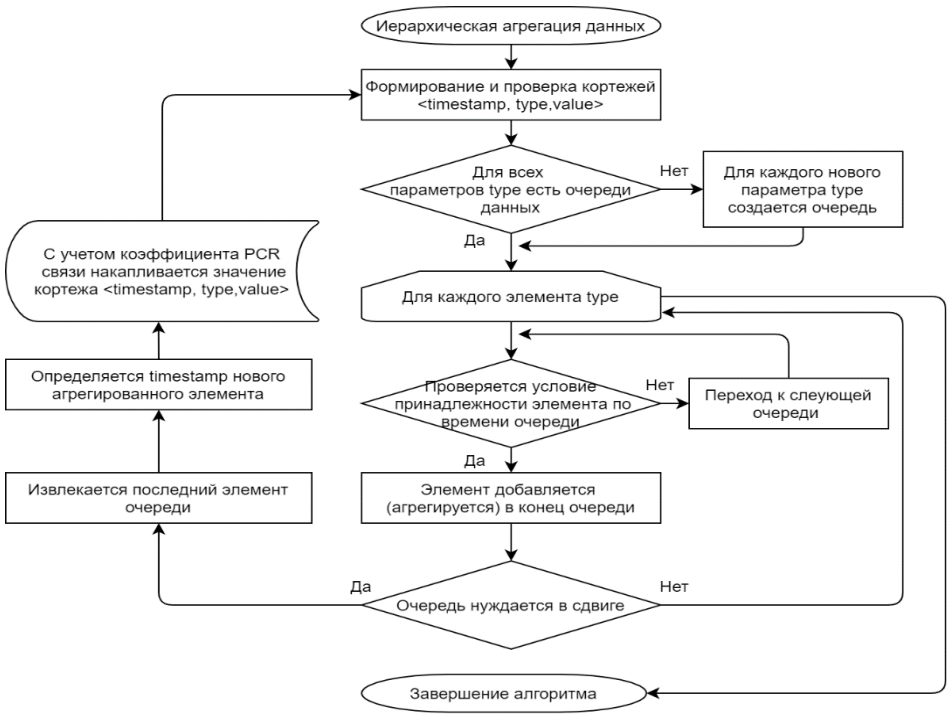

Рис.2. Алгоритм агрегации наборов данных

Приведенный подход к построению модуля сбора и обработки данных позволил построить иерархии временных рядов исходных наборов данных за время порядка ¼ времени обнаружения сетевой атаки. Поставленные перед системой сбора и подготовки данных при мониторинге

безопасности киберфизических систем задачи были решены – для случая с одним типом данных, обрабатываемых в системе (коммуникационные данные или один вид физических данных).

При совместном анализе семантически гетерогенных данных число и объем пересылок информации между узлами анализа требовали все еще существенных накладных расходов. Для решения этой проблемы был предложен переход к многомерному анализу и представлению системы, что потребовало также модификации метода агрегации.

2.4 Обработка данных и многомерная модель системы

Подход на основе совместного анализа разнородных данных при мониторинге безопасности киберфизических систем приводит к концепции многомерного представления объекта защиты на основе множества временных рядов параметров. В области анализа данных существует схожая технология - OLAP (Online Analytic Processing). Основное отличие OLAP – систем это ориентированность на хранимые данные (а не оперативно поступающую потоковую информацию) и, в том числе за счет этого, значительное время для генерации ответа на запрос. Потоковая аналитика, применяющаяся в том числе при мониторинге информационной безопасности киберфизических систем основана на потоковой модели запроса [9]. В этом случае возможна адаптация OLAP – методов для анализа данных мониторинга безопасности. Общая схема обработки данных приведена на рис. 3.

Основные отличия подхода потоковой аналитики в системе мониторинга безопасности КФС от традиционного OLAP заключаются в следующем:

1. Подсистема сбора и подготовки данных, формирующая временные ряды параметров, образующих многомерную модель системы.
2. Учет и поддержание иерархических связей внутри рядов одних и тех же параметров для поддержки применения методов анализа на основе самоподобия и фрактальных характеристик.
3. Представление витрин данных и визуализации данных как срезов данных по заданным заранее, в рамках потоковой модели запроса, характеристикам от различных потребителей данных.

В качестве потребителей данных выступают программные средства и модули осуществляющие непосредственный поиск аномалий, обнаружение вторжений и другие функции безопасности, вне зависимости от специфики используемых в них аналитических методов.

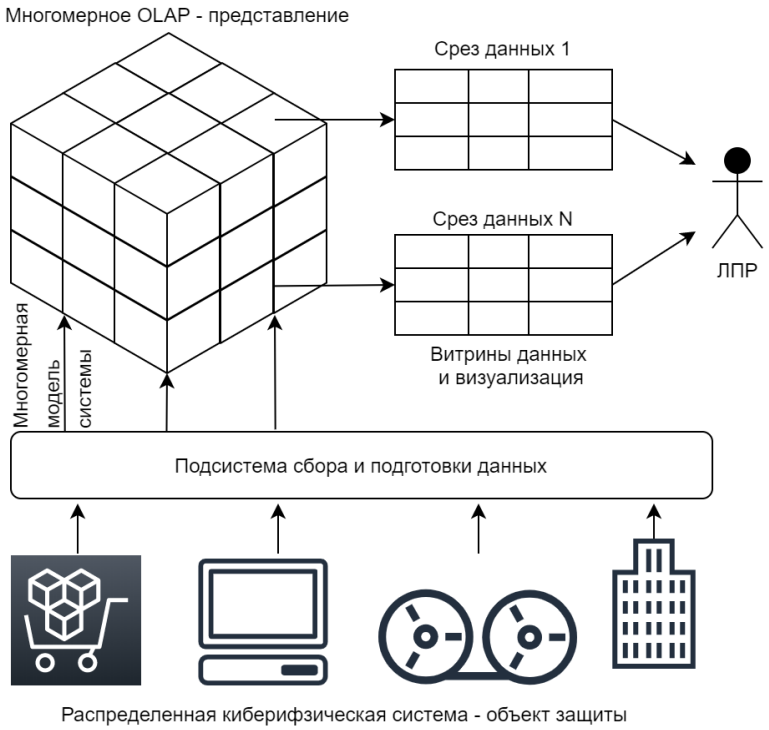

Рис.3. Схема обработки данных на основе многомерной модели системы

Для эффективного функционирования обработки данных на основе многомерной модели необходим механизм организации данных внутри OLAP – куба обеспечивающий не только поддержку иерархических PCR связей между рядами данных, но и отражающий взаимосвязь между различными

наборами параметров. Для этого иерархическая агрегация, алгоритм который для подсистемы мониторинга был приведен выше, должна быть дополнена методом многомерной агрегации.

Для решения этой задачи применяется подход моделирования данных, заключающейся в отделении данных от структуры, и генерация производных наборов «по требованию», на основании задания потокового запроса (рис. 4).

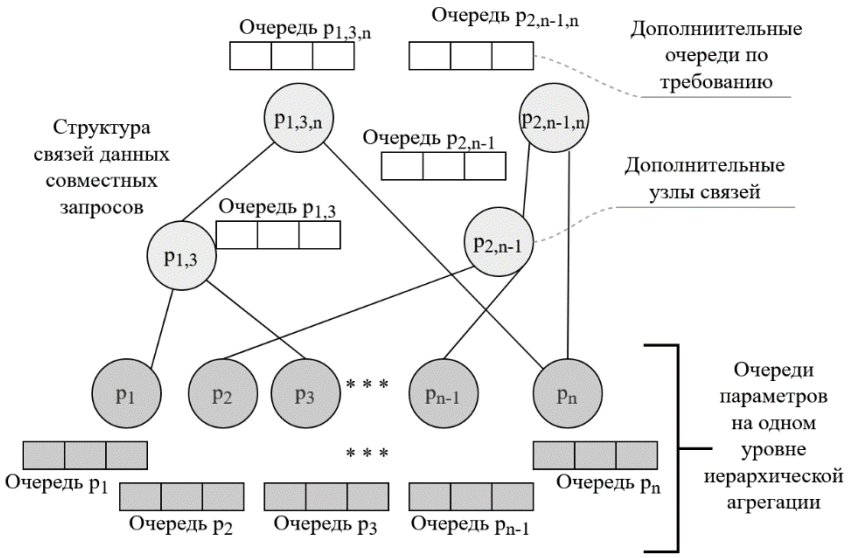

Рис.4. Графовые структуры для многомерной агрегации данных

Каждый потоковый запрос генерирует набор требований вида $\{< \{Time_{Period}, Type\}, F_A\ (Data) >\}$ где $Time_{Period}$ – требуемый временной период, $Type$ – параметр агрегации, $F_A\ (Data)$ – функция совместной агрегации данных . На основании таких наборов генерируется граф агрегации $G(V, E)$, вершины которого представляют собой наборы агрегируемых ключей очередей, а ребра – связи по передаче данных, формируя структуру агрегации. Так как совместный анализ данных осуществляется на базе одних и тех же временных периодов, в условиях самоподобных процессов КФС, по умолчанию при поступлении запроса $< \{Time_{Period}\ , Type\ \}_1, F_{A1}\ (Data) >$ создаются ключевые структуры для всех наборов данных во временных иерархиях. Это обуславливается поддержкой само подобной обработки данных и спецификой объекта мониторинга.

В свою очередь созданные структуры ассоциируются с исходными очередями параметров и связанные с ними производные очереди (см. рис. 4) создаются и заполняются данными по требованию. Такое решение позволяет поддерживать структуру агрегации с учетом само подобия с минимальными накладными расходами.

4 Оценка эффективности предложенных решений

Эффективность предложенных на данном этапе решений для задачи адаптивного мониторинга информационной безопасности оценивалась на основе соответствия требованиям, предъявляемым к такого рода системам и характеристик многомерной агрегации данных, примененной для обеспечения информированности и подготовки данных в отношении множества методов анализа.

Было проведено тестирование разработанной программы на датасете, полученном в результате работы системы по очистке воды. Данные были собраны в результате 11 дней непрерывной работы системы, в течение которых на систему производились различные атаки.

На данном этапе работы в отношении многомерной агрегации оценивались размер и характеристики графа, такие как количество узлов и ребер в графе, количество связанных компонент, количество рёбер и узлов в самой большой связанной компоненте. Эти параметры напрямую влияют на скорость получения доступа к узлам графа и скорость получения всех узлов, связанным с поступившим потоковым запросом. Был проведён анализ зависимости размера графа от количества пар агрегируемых параметров (0 – нет агрегируемых параметров, 1 – одна пара агрегируемых параметров, 2 – две и т.д.). Построены графики зависимости количества узлов и рёбер в графе, а также количества узлов и рёбер в самой большой связной компоненте графа, и количества связных компонент, от количества таких пар.

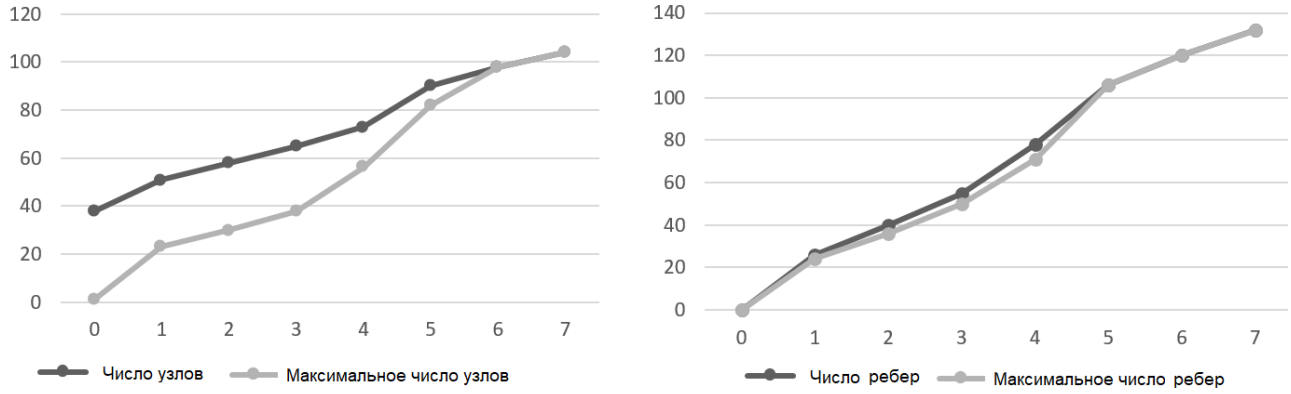

Рис.5. Зависимости числа и максимального числа ребер и вершин графа от количества агрегируемых параметров

На рисунке 5 приведены зависимости количества узлов и рёбер в графе, а также количества узлов и рёбер в самой большой связной компоненте графа от количества агрегируемых параметров для тестовой киберфизической системы. При количестве пар агрегируемых параметров больше 5 практически все узлы графа объединяются в одну связанную компоненту, и эффективность извлечения пар параметров ухудшается.

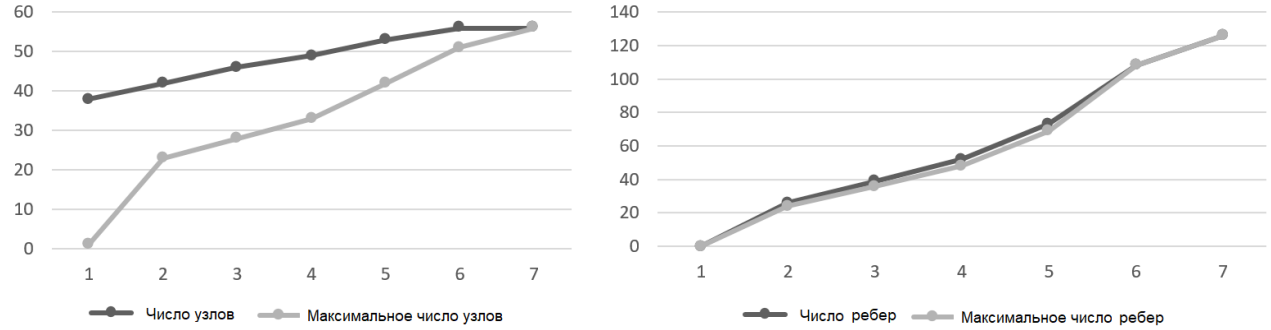

Рис.6. Зависимости числа и максимального числа ребер и вершин графа от глубины агрегации параметров

На рисунке 6 приведены зависимости количества узлов и рёбер в графе, а также количества узлов и рёбер в самой большой связной компоненте графа от глубины агрегации параметров для тестовой киберфизической системы. При глубине вложенности агрегируемых параметров больше 6 практически все узлы графа объединяются в одну связанную компоненту, и эффективность извлечения пар параметров ухудшается.

В результате оценки можно сделать вывод об эффективности предложенного метода многомерной агрегации и системы сбора, подготовки и обработки данных в целом при глубине агрегации и числе наборов агрегируемых параметров порядка 4-х. На сегодняшний день это достаточное число пар параметров и глубина вложенности агрегации для комбинированных методов анализа при мониторинге безопасности КФС. При этом размер данных для анализа в совокупности представляет собой порядка нескольких мегабайт (порядка половины мегабайта для 4-х пар агрегаций и/или глубины вложенности 4) что позволяет при линейном росте размерности графа (см. рис. 5 и 6) говорить о размещении всего объема данных для анализа в оперативной памяти одного узла. Линейный рост размерности говорит о линейном возрастании нагрузки на систему что, с учетом линейного роста размера данных и текущих значений размерности (несколько мегабайт) говорит о хорошем запасе производительности для большего числа агрегаций возрастающей глубины анализа само подобия.

Заключение

В результате работы были выявлены основные особенности мониторинга безопасности распределенных киберфизических систем. Среди них сочетание коммуникационных и конечных данных при анализе, динамическое изменение среды и аналитических методов, периодичность и само подобие протекающих процессов (как коммуникационных в виде сетевого трафика, так и физических). Такие особенности обуславливают необходимость адаптивности системы мониторинга, в том числе в

отношении динамического измерения методов анализа данных, а также поддержку особенностей анализа самоподобных данных при высокой нагрузке. Высокая нагрузка на мониторинга безопасности современных КФС обусловлена ростом связанных устройств, распределенностью и масштабом систем и, как следствие, ростом скорости и объемов поступающей информации.

Для решения этой задачи совместно с традиционными методами конвейерной и параллельной обработки данных в системе мониторинга предлагаются специфические методы агрегации данных. Иерархическая агрегация применительно к адаптивному мониторингу безопасности обеспечивает поддержку анализа потоковых само подобных данных минимизируя объем оперативных данных поступающих алгоритму – потребителю и число их пересылок между узлами.

Для обеспечения динамичности в отношении методов анализа предлагается применение многомерной модели системы в виде наборов иерархий временных рядов, связанных отношением агрегации на основании графа агрегации. Тестирование было проведено на наборе данных, полученном в результате работы системы по очистке воды. Данные были собраны в результате 11 дней непрерывной работы системы, в течение которых на систему производились различные атаки. Показано, что хорошие рабочие характеристики предложенным методом можно достичь при агрегации порядка 4-х пар агрегируемых параметров и сходной глубине агрегации. Абсолютные значения объема данных при таких характеристиках и линейный рост размерности графа говорят о хорошем потенциале масштабируемости при росте числа параметров и/или глубины агрегации.

Литература

1. Актуальные киберугрозы: итоги 2019 года https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2019/#id16
2. *Rostuntsova A. A., Ryskin N. M., Ginzburg N. S.* Self-Similar Analysis of Short Pulse Amplification and Generation in Cherenkov-type Devices // 2019 International Vacuum Electronics Conference (IVEC), Busan, Korea (South). 2019. - pp. 1-2. DOI: 10.1109/IVEC.2019.8744788.
3. *Coletta A., Armando A.* Security Monitoring for Industrial Control Systems // Security of Industrial Control Systems and Cyber Physical Systems. CyberICS 2015, WOS-CPS 2015. – LNCS, Springer, 2015. – Vol. 9588. – P. 48–62.
4. *Brandauer C., Dorfinger P., Paiva P. Y. A.* Towards scalable and adaptable security monitoring // IEEE 36th International Performance Computing and Communications Conference (IPCCC), San Diego, CA, 2017. - pp. 1-6. DOI: 10.1109/PCCC.2017.8280502.
5. *Kalinin, M.O., Lavrova, D.S., Yarmak, A.V.* Detection of Threats in Cyberphysical Systems Based on Deep Learning Methods Using Multidimensional Time Series. // Aut. Control Comp. Sci. 52, 912–917 (2018). DOI: 10.3103/S0146411618080151
6. *Печенкин А.И., Полтавцева М.А., Лаврова Д.С.* An Approach to Data Normalization in the Internet of Things for Security Analysis // Программные продукты и системы. 2016. № 2. - С. 83-88.
7. *Полтавцева М.А., Лаврова Д.С., Печенкин А.И.* Планирование задач агрегации и нормализации данных интернета вещей для обработки на многопроцессорном кластере // Проблемы информационной безопасности. Компьютерные системы. 2016. № 1. С. 37-46.
8. *Poltavtseva M.A., Zegzhda P.D., Pankov I.D.* The Hierarchial Data Aggregation Method in Backbone Traffic Streaming Analyzing to Ensure Digital Systems Information Security // Proceedings of 2018 11th International Conference "Management of Large-Scale System Development". MLSD 2018. 2018. - С. 8551916.
9. *Kleppmann M.* Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems. – Boston: O'Reilly Media, 2017. – 640p.