ВЫЯВЛЕНИЕ ОТКЛОНЕНИЙ В ПОВЕДЕНЧЕСКИХ ПАТТЕРНАХ ПОЛЬЗОВАТЕЛЕЙ КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ ТОПОЛОГИЧЕСКИХ ПРИЗНАКОВ

Нашивочников Н.В.[1]

Цель статьи: *в работе рассматривается применение топологического анализа данных для выявления и классификации паттернов поведения пользователей корпоративных информационных ресурсов в задачах обеспечения кибербезопасности.*

Метод: *исследование основано на математических концепциях теории персистентных гомологий (симплициальные комплексы, диаграммы персистентности, функции фильтрации, топологические дескрипторы), теоретических моделях метрических пространств.*

Полученный результат: *дается формализованное определение топологических признаков, раскрываются свойства и связь кривых Бетти с диаграммами персистентности и другими известными топологическими дескрипторами, применяемыми в топологическом анализе данных. Результаты практической апробации на представленных в виде временных рядов данных из системы мониторинга работы пользователей корпоративной сети с корпоративными информационными ресурсами, подтверждают, что использование положительно определенных ступенчатых функций для построения кривых Бетти обеспечивает приемлемую вычислительную сложность процедур, необходимых для определения и классификации поведенческих паттернов пользователей. Предложенный на основе топологических дескрипторов и модифицированных функций желательности Харрингтона подход позволяет надежно фиксировать отклонение активностей пользователя от типового паттерна поведения, что потенциально может представлять собой инцидент безопасности.*

Научная новизна: *заключается в применении топологического анализа данных с использованием кривых Бетти для выявления и классификации паттернов поведения пользователей корпоративных информационных ресурсов.*

Ключевые слова: *кибербезопасность, UEBA, топологический анализ данных, кривые Бетти, временные ряды.*

DOI:10.21681/2311-3456-2023-4-12-22

1. Введение

Для выявления отклонений в поведенческих паттернах (типовых шаблонах действий) пользователей корпоративных информационных ресурсов, которые потенциально могут представлять собой инциденты безопасности, в UEBA[2] наряду со статистическими методами находят применение методы расширенной аналитики[3], включая алгоритмы глубокого обучения [1,2]. Известно, что при машинном обучении большой класс задач можно разделить на задачи обучения с учителем и без учителя [3]. Для задач обучения с учителем выходные показатели предсказываются (объясняются) на основе входных показателей. Задачи же обучения без учителя можно рассматривать как извлечение признаков из данных, предполагая, что последние сопровождаются неструктурированным шумом. Многие методы в науке о данных можно квалифицировать именно как методы обучения без учителя [3]. К ним, например, относятся методы ассоциации, методы кластеризации (линейные и нелинейные), методы уменьшения размерности и матричная факторизация, топологический анализ данных и это лишь некоторые из них.

2 User and Entity Behavior Analytics // Энциклопедия «Касперского». https://encyclopedia.kaspersky.ru/glossary/ueba/.

3 Advanced Analytics // Gartner IT Glossary. URL: https://www.gartner.com/it-glossary/advanced-analytics/.

1 Нашивочников Николай Васильевич, CISSP, заместитель генерального директора – технический директор, ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: cto@gaz-is.ru.

Развиваемый в последнее время топологический анализ данных представляет собой достаточно новую область в науке о данных, целью которого является раскрытие, понимание и использование топологической и геометрической структуры, содержащейся в данных [4, 5]. Причем для получения количественных оценок этих структур, как правило, используются вычисляемые специальным образом так называемые топологические дескрипторы [5], к которым уже можно применять широкий арсенал количественных методов анализа, базирующихся на использовании тех или иных метрик [6].

В статье рассмотрены такие топологические дескрипторы как кривые Бетти, которые определенным образом обобщают диаграммы персистентности (дескрипторы признаков в топологическом анализе данных). Их можно вычислять за линейное время. Эти дескрипторы используются для выявления отклонений в поведенческих паттернах пользователей корпоративных информационных ресурсов. Выявленные отклонения фиксируются UEBA как потенциальный инцидент безопасности. В отличие от работы [7], посвященной задаче разработки общей методики использования топологического анализа данных в системах поведенческой аналитики, предметом настоящей работы является подход для выявления отклонений в поведенческих паттернах пользователей корпоративных информационных ресурсов с использованием кривых Бетти.

2. Диаграммы персистентности, кривые персистентности и кривые Бетти

Центральным элементом анализа данных, основанного на теории персистентных гомологий, является диаграмма персистентности [4,5,7,14], которая по существу хранит связанные критические точки функции фильтрации. Поясним, что под этим понимается. Рассмотрим симплициальный комплекс $\mathcal{K}$[4] и определенную на нем вещественнозначную функцию $f:\mathcal{K}\to\mathbb{R}$. Функция f называется симплексно-монотонной, если для любого симплекса $\sigma'\subseteq\sigma$ выполняется неравенство $f(\sigma')\le f(\sigma)$. Отметим, что если функция f является симплексно – монотонной, то множества ее подуровней $f^{-1}(-\infty,a]$ являются подкомплексами симплициального комплекса $\mathcal{K}$ для любого $a\in\mathbb{R}$. Полагая

$$\mathcal{K}_i=f^{-1}(-\infty,a_i];\ i=1,\dots,n;\ a_0=-\infty,$$

получаем фильтрацию

$$F:\varnothing=\mathcal{K}_0\hookrightarrow\mathcal{K}_1\hookrightarrow\dots\mathcal{K}_i\hookrightarrow\mathcal{K}_{i+1}\dots\hookrightarrow\mathcal{K}_n=\mathcal{K}.$$

Определим вещественнозначную функцию $f:V(\mathcal{K})\to\mathbb{R}$ на множестве вершин $V(\mathcal{K})$ комплекса $\mathcal{K}$[5] и построим симплициальную фильтрацию, которую обозначим как F_f. Персистентная диаграмма $\mathrm{Dgm}_p(F_f)$[6], как множество точек расширенной плоскости $\overline{\mathbb{R}}^2$ [4,5] интегрирует в себя определенную топологическую информацию из симплициального комплекса $\mathcal{K}$ относительно функции фильтрации f, индуцирующей фильтрацию F_f[7]. Однако на практике эта информация может потерять какой-либо смысл, если небольшая вариация функции фильтрации f приведет к резкому изменению диаграммы $\mathrm{Dgm}_p(F_f)$. Следует заметить, что функция f редко бывает известна точно. Обычно известно ее некоторое приближение – $\tilde{f}$. Тогда, если диаграмму $\mathrm{Dgm}_p(F_f)$ можно аппроксимировать диаграммой $\mathrm{Dgm}_p(F_{\tilde{f}})$ с приемлемой точностью, то вся топологическая информация, извлекаемая из симплициального комплекса $\mathcal{K}$, по-прежнему останется актуальной и диаграмма $\mathrm{Dgm}_p(F_{\tilde{f}})$ будет служить источником информации о фильтрации F_f. Ситуация упрощается в связи с тем, что диаграммы персистентности устойчивы [8].

Пусть $\mathrm{Dgm}_p(F_f)$ и $\mathrm{Dgm}_p(F_g)$ – диаграммы персистентности для функций фильтрации f и g[8], соответственно, π – биекция между точками из $\mathrm{Dgm}_p(F_f)$ и $\mathrm{Dgm}_p(F_g)$, $\pi:\mathrm{Dgm}_p(F_f)\to\mathrm{Dgm}_p(F_g)$.

Определение (Расстояние узкого места)

Пусть $\Pi_\pi=\left\{\pi\middle|\mathrm{Dgm}_p(F_f)\to\mathrm{Dgm}_p(F_g)\right\}$ – множество всех биекций точек из диаграммы $\mathrm{Dgm}_p(F_f)$ в диаграмму $\mathrm{Dgm}_p(F_g)$. Рассмотрим расстояние между двумя точками $x=(x_1,x_2)$ и $y=(y_1,y_2)$ в L_∞ – норме:

$$\|x-y_\infty\|=\max\{|x_1-x_2|,|y_1-y_2|\},$$

считая, что $\infty-\infty=0$.

Расстояние узкого места $\mathrm{d}_b(\bullet,\bullet)$ между двумя диаграммами определяется выражением вида [8,14]:

$$\mathrm{d}_b\left(\mathrm{Dgm}_p(F_f),\mathrm{Dgm}_p(F_g)\right)=\inf_{\pi\in\Pi_\pi}\sup_{x\in Dgm_p(F_f)}\|x-\pi(x)\|_\infty.$$

4 Технология построения симплициальных комплексов по наборам данных изложена в работах [4,7,8].

5 Такая функция называется вершинной.

6 Здесь p = 0 признак гомологического класса. Если p=0, то имеется в виду компонента связанности. При p=1 имеется в виду «отверстие, дыра» и т.д., p≥0,1,2,... , .

7 Далее рассматривается только фильтрация конечных комплексов, которая, в свою очередь, делает конечной и каждую группу гомологий.

8 Диаграммы $\mathrm{Dgm}_p(F_f)$ и $\mathrm{Dgm}_p(F_g)$ могут иметь разную мощность для недиагональных точек. В связи с этим они включают и точки на диагонали, каждая из которых имеет бесконечную кратность. Это позволяет «заимствовать» данные точки, когда это необходимо для корректного определения биекций [8,15].

Утверждение [8]

1. Расстояние узкого места $d_b(\bullet,\bullet)$ является метрикой в пространстве диаграмм персистентности.

2. Метрика $\mathrm{d}_b(X,Y)=0$ тогда и только тогда, когда $X=Y$. Более того,

$$\mathrm{d}_b(X,Y)=\mathrm{d}_b(Y,X);$$
$$\mathrm{d}_b(X,Y)\le\mathrm{d}_b(X,Z)+\mathrm{d}_b(Z,Y).$$

Если расстояние узкого места d_b принимается и как расстояние в пространстве гомологических модулей $\mathcal{H}_pF_f$, то есть, если $\mathrm{d}_b\left(\mathcal{H}_pF_f,\mathcal{H}_gF_g\right)=\mathrm{d}_b\left(\mathrm{Dgm}_p\left(F_f\right),\mathrm{Dgm}_p\left(F_g\right)\right)$, то так определенное расстояние может и не являться метрикой. Действительно, первая аксиома метрики не выполняется, если гомологическим модулям разрешено создавать и уничтожать гомологические классы при одних и тех же значениях функций фильтрации. Эти классы с нулевой персистентностью порождают точки на диагонали диаграммы [15,16]. Поскольку точки на диагонали имеют бесконечную кратность, то два модуля, различающиеся числом таких классов с нулевой персистентностью, могут иметь диаграммы с нулевым расстоянием узкого места. Если такие случаи допускаются, то расстояние узкого места d_b становится «псевдометрикой» в пространстве гомологических модулей. Это означает, что она удовлетворяет всем аксиомам метрики, кроме первой (аксиомы тождества).

Следующие теоремы дают количественную оценку понятия устойчивости диаграммы персистентности. Есть две версии: одна сформулирована и доказана для случая симплициальной фильтрации, а другая — для случая пространственной фильтрации [5,8,15,16]. Дадим их формулировки для обоих случаев. Для двух функций $f,g:X\to\mathbb{R}$ метрика в L_∞ определяется как

$$\|f-g\|_\infty:=\sup_{x\in X}|f(x)-g(x)|.$$

Теорема устойчивости для симплициальных фильтраций

Пусть $f,g:K\to\mathbb{R}$ — две симплексно-монотонные функции, порождающие две симплициальные фильтрации F_f и F_g. Тогда для любого $p\in\mathbb{N}$,

$$\mathrm{d}_b\left(\mathrm{Dgm}_p\left(F_f\right),\mathrm{Dgm}_p\left(F_g\right)\right)\le\|f-g\|_\infty.$$

Для второй версии теоремы устойчивости требуется, чтобы функции $f,g:K\to\mathbb{R}$, упомянутые в теореме, были «хорошими» в том смысле, что они являются «ручными»[9].

Теорема устойчивости для пространственных фильтраций

Пусть X — триангулируемое топологическое пространство, а $f,g:K\to\mathbb{R}$ — две ручные функции, порождающие две пространственные фильтрации F_f и F_g, где множества подуровней включают критические точки $\{a_i\}_{i=1}^n$. Тогда для каждого $p\in\mathbb{N}$

$$\mathrm{d}_b\left(\mathrm{Dgm}_p\left(F_f\right),\mathrm{Dgm}_p\left(F_g\right)\right)\le\|f-g\|_\infty.$$

Существует еще одна метрика, называемая q — расстоянием Вассерштейна, с помощью которого можно также сравнивать диаграммы персистентности.

Определение (расстояние Вассерштейна)

Пусть $\Pi_\pi=\left\{\pi:\mathrm{Dgm}_p\left(F_f\right)\to\mathrm{Dgm}_p\left(F_g\right)\right\}$ — множество всех биекций точек из диаграммы $\mathrm{Dgm}_p\left(F_f\right)$ в диаграмму $\mathrm{Dgm}_p\left(F_g\right)$. Для любых $p,q\in\mathbb{N},q\ge1$ q — расстояние Вассерштейна $\mathrm{d}_{W,q}$ определяется как

$$\mathrm{d}_{W,q}\left(\mathrm{Dgm}_p\left(F_f\right),\mathrm{Dgm}_p\left(F_g\right)\right)=$$
$$=\inf_{\pi\in\Pi_\pi}\sum_{x\in\mathrm{Dgm}_p\left(F_f\right)}\left(\|x-\pi(x)\|_\infty^q\right)^{1/q}.$$

Расстояние $\mathrm{d}_{W,q}$, как и расстояние узкого места, является метрикой в пространстве диаграмм персистентности. Оно также обладает свойством устойчивости, хотя и не таким сильным, как в теореме устойчивости для пространственных фильтраций.

Это свойство можно сформулировать в виде следующего утверждения.

Утверждение (Свойство устойчивости для метрики Вассерштейна)

Пусть $f,g:K\to\mathbb{R}$ — две липшицевы функции, определенные на триангулируемом компактном метрическом пространстве X. Тогда существуют константы $C\in\mathbb{R}_+^1,C>0$ и $k\in\mathbb{N}$ такие, что для любого $p\in\mathbb{N}$, $p\ge k$

$$\mathrm{d}_{W,q}\left(\mathrm{Dgm}_p\left(F_f\right),\mathrm{Dgm}_p\left(F_g\right)\right)\le C\|f-g\|_\infty^{1-\frac{k}{p}}.$$

Расстояние узкого места можно вычислить с по-

9 Напомним, что функция $f:X\to\mathbb{R}$ называется ручной, если у нее существует лишь конечное число гомологических критических значений $\{a_i\}_{i=1}^n$. Другими словами, если группы гомологий ее множеств подуровней имеют конечные ранги и эти ранги изменяются в критических точках $\{a_i\}_{i=1}^n$, которых конечное число.

мощью алгоритма поиска совершенных паросочетаний[10] в двудольных графах, а расстояние Вассерштейна — с помощью, например, алгоритма поиска совершенных паросочетаний минимального веса во взвешенных двудольных графах.

Теоремы устойчивости гарантируют, что расчеты будут устойчивы к возмущениям и неизбежному возникновению шума в реальных данных. Это достигается за счет очень большого времени выполнения расчетов расстояний между диаграммами персистентности: для обеих метрик они имеют сложность не менее $O\left(n^{2.5}\right)$ или, в наивной реализации, $O\left(n^3\right)$. Используя рандомизированные алгоритмы, можно достичь сложности $O\left(n^\omega\right)$, где $\omega < 2,38$ (данный параметр определяет оценку наилучшего времени матричного умножения). Дальнейшее снижение вычислительной сложности возможно, если обеспечено максимальное приближение к «правильной» метрике [5,13,14,17-20]. Тем не менее, и эти алгоритмы достаточно сложно реализовать непосредственно без использования специальных алгоритмов редукции, а их производительность хуже, чем $O\left(n^2\right)$. Последнее означает, что для сравнения больших наборов диаграмм персистентности целесообразно выбирать иные дескрипторы, которые бы, с одной стороны отражали сакраментальную суть диаграмм, а с другой стороны, существенно улучшали производительность расчетов. Таким дескриптором является кривая Бетти, которую можно легко и быстро вычислить. Если говорить о задачах поведенческой аналитики, то данные дескрипторы были применены в работе [7] для формирования профилей поведения пользователей информационных ресурсов и выявления текущих отклонений от них. В настоящей работе приводится несколько иное описание кривых Бетти, базирующееся на использовании положительно определенных ступенчатых функций.

В некотором смысле предшественником кривых Бетти можно считать кривые персистенции, которые часто используются при анализе комплексов Морса–Смейла [9]. Поясним, что понимается под кривой персистентности. Рассмотрим двумерное компактное гладкое многообразие без края $\mathbb{M}$ и предположим, что на нем задана некоторая гладкая функция $f : \mathbb{M} \to \mathbb{R}$. Кроме того, будем считать, что на данном многообразии заданы риманова метрика и ортонормированная локальная система координат в произвольной точке a данного многообразия. Как следует из леммы Морса [9,10] градиент функции f образует гладкое векторное поле на многообразии $\mathbb{M}$ с нулями в критических точках[11]. Примерами критических точек являются максимумы (функция f уменьшается во всех направлениях), минимумы (функция f увеличивается во всех направлениях) и седла (функция f переключается между уменьшением и увеличением четыре раза вокруг точки). Используя местные координаты произвольной точки на многообразии $\mathbb{M}$, вычислим гессиан функции f. Критическая точка является невырожденной, когда гессиан отличен от нуля[12]. Тогда в соответствии с леммой Морса в окрестности любой невырожденной критической точки $a \in \mathbb{M}$ можно построить такую локальную систему координат, что функция f представима в виде $f\left(x_1, x_2\right) = f\left(a\right) \pm x_1^2 \pm x_2^2$. Количество знаков минус является индексом точки a и различает различные типы критических точек: минимумы имеют индекс 0, седла имеют индекс 1, а максимумы имеют индекс 2. Заметим, что функция f является функцией Морса тогда и только тогда, когда все ее критические точки невырождены и имеют попарно различные значения. В любой регулярной точке многообразия $\mathbb{M}$ существует ненулевой градиент функции f. Тогда «двигаясь» по этому вектору от точки к точке можно «нарисовать» на многообразии интегральную линию, которая начинается в одной критической точке и заканчивается в другой критической точке, хотя технически (формально) не содержит ни одну из них. Поскольку интегральные линии монотонно восходят[13] (критические точки могут быть только точками минимума, максимума и седловыми точками), то две конечные точки не могут совпадать. Так как функция f по определению – гладкая, то две такие интегральные линии либо не пересекаются, либо совпадают. Множество интегральных линий покрывает все многообразие, кроме критических точек. Таким образом, интегральные линии, исходящие из критической точки a – это множество точек, которые «текут» из a. Такие интегральные линии называют «кривыми персистентности». Формально такая кривая подсчитывает количество критических точек, которые возникают при данном уровне параметра масштаба фильтрации $\varepsilon \in \mathbb{R}_+$ (текущем уровне функции f). Кроме того, она используется для определения соответствующего порога (уровня функции f) или, иногда

10 Паросочетанием в двудольном графе называется любое множество попарно несмежных ребер (у них нет общих вершин).

11 Фоменко А. Т., Фукс Д. Б. Курс гомотопической топологии. — М.: Наука; ГРФМЛ, 1989. — 528 с.

12 Это свойство не зависит от выбора системы координат.

13 Это справедливо, если f – возрастающая функция (функция высоты). Данное допущение не является принципиальным в силу произвольности выбора функции f (ее можно выбрать и убывающей).

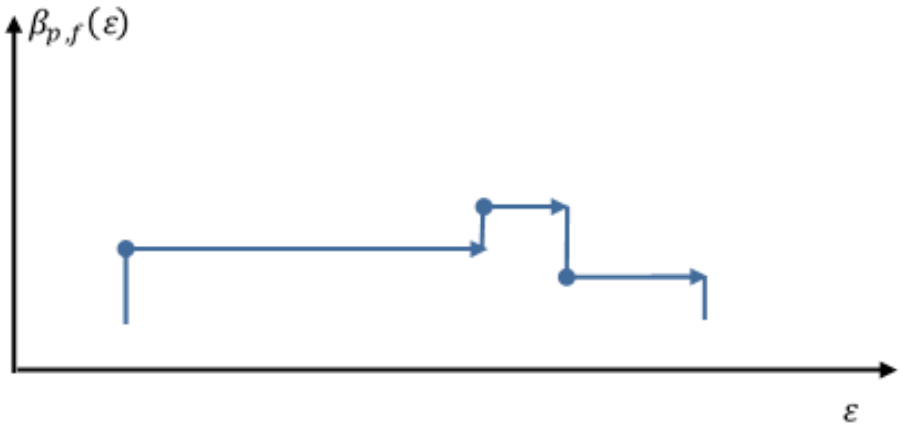

Рис.1. К определению функции $\beta_{p,f}(\varepsilon)$

говорят, параметра «отсечки» для оптимизации числа критических точек функции f [14]. В приложениях требуется численная мера «важности» критических точек, которую можно использовать, например, для упрощения комплекса Морса-Смейла. Для этого критические точки объединяются в пары и подсчитывается абсолютная величина разницы между их высотами как мера «важности»[15]. Чтобы построить такие пары критических точек представим, что двумерное многообразие $\mathbb{M}$ перемещается в направлении увеличения высоты (например, функция f представляет собой функцию высоты). Это эквивалентно сортировке вершин по высоте и поэтапному построению триангуляции $\mathcal{K}$ многообразия $\mathbb{M}$ по одной нижней звезде за шаг. Топология частичной триангуляции меняется всякий раз, когда добавляется критическая вершина, и остается неизменной, когда добавляется обычная вершина. За исключением некоторых случаев, связанных с типом многообразия $\mathbb{M}$, каждое такое изменение либо создает компонент связанности или кольцо (структуру с отверстием), либо разрушает компонент связанности (путем слияния двух компонент) или кольцо (путем заполнения отверстия). Соединим вершину v, которая разрушает структуру триангуляции (компонент или кольцо), с вершиной u, создавшей ту структуру триангуляции, что вершина v разрушила. Тогда персистентность (обозначается как «pers») можно интерпретировать как «задержку» между событиями разрушения (вершина v) и создания (вершина u) структур в триангуляции, т.е.:

$$\mathrm{pers} = f(v) - f(u).$$

Мультимножество таких пар, специальным образом отображенное на плоскость, представляет собой диаграмму персистентности. Таким образом, легко видеть, что кривые персистентности имеют опосредованную связь с диаграммами персистентности.

Рассмотрим теперь симплициальную фильтрацию F_f[16], связанную с некоторым облаком данных M, и будем следить за изменением рангов групп гомологий (чисел Бетти – $\beta_{p,f}, p = 0,1,\ldots,$) в процессе фильтрации, т.е. построим зависимости $\beta_{p,f} = \beta_{p,f}(\varepsilon)$. Они представляют собой кусочно-постоянные функции[17], которые изменяются скачкообразно только в критических точках. Типичный вид зависимости $\beta_p = \beta_{p,f}(\varepsilon)$ для некоторого фиксированного $p \in \mathbb{N}_+$ представлен на рис.1.

Эта зависимость, по аналогии с результатами работы [7], называется «кривой Бетти». Поскольку функция $\varepsilon \mapsto \beta_{p,f}(\cdot) : \mathbb{R} \to \mathbb{N}$ (см. рис. 1) является ступенчатой, то она для любого p представима в виде:

$$\beta_{p,f}(\varepsilon) = \sum_{i=1}^{N} \alpha_i^{(p,f)} \mathbb{I}_{\Delta_i}(\varepsilon), \qquad (1)$$

где $\alpha_i^{(p,f)}$ – количество $p-x$ гомологических классов в гомологическом модуле, возникающих при фильтрации F_f топологического пространства, сопоставляемого рассматриваемому набору данных[18] и отвечающих i – му промежутку разбиения $\Delta_i = [\varepsilon_i, \varepsilon_{i+1}]$, $\varepsilon_i, \varepsilon_{i+1} \in \mathbb{R}_+$ диапазона изменения параметра масштаба фильтрации $\varepsilon \in \mathbb{R}_+$;

$\bigcup_{i=1}^{N} \Delta_i \subseteq [\varepsilon_0, \varepsilon_k]; [\varepsilon_0, \varepsilon_k] \subset \mathbb{R}_+$ – общий диапазон

14 При этом не существует единого стандарта для описания этих кривых.

15 Это можно сделать, поскольку предполагается, что на многообразии $\mathbb{M}$ задана риманова метрика.

16 Не ограничивая общности, полагаем функцию f — «ручной».

17 Точнее, ступенчатые функции, поскольку в силу построения принимают не более чем счетное количество разных значений.

18 Здесь $0 \le \alpha_i^{(p,f)} < \infty$ в силу того, что рассматривается конечномерная фильтрация.

изменения масштаба ε; $\mathbb{I}_{\Delta_i}(\cdot)$ – индикаторная функция.

Приведем ряд свойств функции $\beta_{p,f}(\varepsilon)$. Во-первых, множество функций $\beta_{p,f}(\varepsilon)$ (в дальнейшем будем обозначать его как L_{β_p}) со стандартными операциями поточечного сложения и умножения функций на числа из поля скаляров образуют линейное пространство. Во-вторых, их произведение, частное от деления (если функция-делитель отлична от нуля) и модуль – $\left|\beta_{p,f}(\varepsilon)\right|$ являются также ступенчатыми функциями[19]. При этом сложение двух различных функций $\beta_{p,f}(\varepsilon)$ и $\beta_{p,g}(\varepsilon)$ соответствует объединению диаграмм $\mathrm{Dgm}_p(F_f)$ и $\mathrm{Dgm}_p(F_g)$, отвечающих двум фильтрациям F_f и F_g. Следовательно, если имеется набор из K диаграмм персистентности $\left\{\mathrm{Dgm}_p(F_{f_i})\right\}_{i=1}^{K}$, то, используя соотношение (1) и тот факт, что ступенчатые функции образуют линейное пространство, легко вычислить среднюю по набору функцию $\bar{\beta}_{p,f}(\cdot)$[20]:

$$\bar{\beta}_{p,f}(\cdot)=\frac{1}{K}\sum_{i=1}^{K}\beta_{p,f_i}(\cdot), \tag{2}$$

которая также будет ступенчатой функций.

Для ступенчатых функций интеграл Лебега определен однозначно, а так как $\left|\beta_{p,f}(\cdot)\right|$ также представляет собой ступенчатую функцию, то существует и интеграл $\int_{\mathbb{R}}\left|\beta_{p,f}(x)\right|dx$. Определим на множестве L_{β_p} функционал:

$$\left\|\beta_{p,f}\right\|_{L_{\beta_p}}=\int_{\mathbb{R}_+}\left|\beta_{p,f}(x)\right|dx. \tag{3}$$

Функционал (3) в общем случае не является нормой на пространстве ступенчатых функций[21]. Для того, чтобы функционал (3) являлся нормой, необходимо выполнение условия $\left\|\beta_{p,f}\right\|_{L_{\beta_p}}>0$, если $\beta_{p,f}\neq 0$. Данное условие выполняется, если считать, что эквивалентные функции[22] из пространства L_{β_p} не «различаются», а принимаются за один и тот же элемент данного пространства. Другими словами, пространство L_{β_p} представляет собой пространство L_1, элементами которого служат классы эквивалентных ступенчатых положительно определенных функций. При этом результатом сложения двух классов будет являться класс, содержащий сумму выбранных представителей из каждого класса [11]. Подчеркнем, что результат сложения этих представителей также является положительно определенной ступенчатой функцией и не зависит от выбранных представителей классов.

Из (3) и определения функций $\beta_{p,f}(\cdot)$ следует, что для каждого $p\in\mathbb{N}_+$ и «ручной» функции фильтрации f выполняется неравенство $\left\|\beta_{p,f}\right\|_{L_{\beta_p}}<\infty$. С другой стороны, подставляя в (3) выражение (1), получаем:

$$\left\|\beta_{p,f}\right\|_{L_{\beta_p}}=\sum_{i=1}^{N}\alpha_i^{(p,f)}\int_{\mathbb{R}_+}\left|\chi_{\Delta_i}(x)\right|dx=\sum_{i=1}^{N}\alpha_i^{(p,f)}\mu(\Delta_i)<\infty, \tag{4}$$

где $\mu(\Delta_i)$ – мера (длина) промежутка $\Delta_i\overset{\text{def}}{\equiv}\left|\varepsilon_i,\varepsilon_{i+1}\right|$, N – число промежутков, на которых число гомологических классов не меняется.

Таким образом, $\left\|\beta_{p,f}\right\|_{L_{\beta_p}}<\infty$ определяет общую персистентность симплициальной фильтрации F_f и, в силу ее ограниченности, выбранный топологический дескриптор (кривая Бетти) будет устойчив относительно малых возмущений.

Метрика на пространстве L_{β_p} вводится следующим образом:

$$\mathrm{d}_{L_{\beta_p}}\left(\beta_{p,f}(\varepsilon),\beta_{p,g}(\varepsilon)\right)= \tag{5}$$
$$=\int_{\mathbb{R}_+}\left|\beta_{p,f}(x)-\beta_{p,g}(x)\right|dx,\ p\geq 0,1,\ldots.$$

Из выражений (4), (5) следует, что вычислительная сложность вычисления метрики $\mathrm{d}_{L_{\beta_p}}(\cdot,\cdot)$ линейна по количеству подинтервалов области определения кривых Бетти.

3. Экспериментальные результаты

В настоящее время по-прежнему является актуальной задача анализа поведения пользователей корпоративной сети при работе с корпоративными информационными ресурсами, работающими зачастую в дистанционном режиме. Общий подход к решению данной задачи с использованием топологических дескрипторов и функций желательности изложен в работе [7]. Ниже приведены результаты исследований с использованием топологических дескрипторов на базе определенных в п.2 кривых Бетти, представленных в виде ступенчатых функций. В качестве анализируемых данных выбраны данные из системы оперативного мониторинга, представляющие собой время активной работы пользователя с корпоративными информационными ресурсами. Эти данные генерируются системой оперативного мониторинга в виде временных рядов таких показателей как активное время

19 Эти свойства вытекают непосредственно из свойств ступенчатых функций.

20 По функции $\bar{\beta}_{p,f}(\cdot)$ достаточно просто построить и «среднюю» диаграмму персистентности $\mathrm{Dgm}_p(F_{\bar{f}})$

21 Введение в функциональный анализ: учеб. пособие / А. С. Кутузов. – Москва; Берлин: Директ-Медиа, 2020. – 481 с.

22 Функции f и g из пространства L_{β_p} считаем эквивалентными, если $f-g=0$.

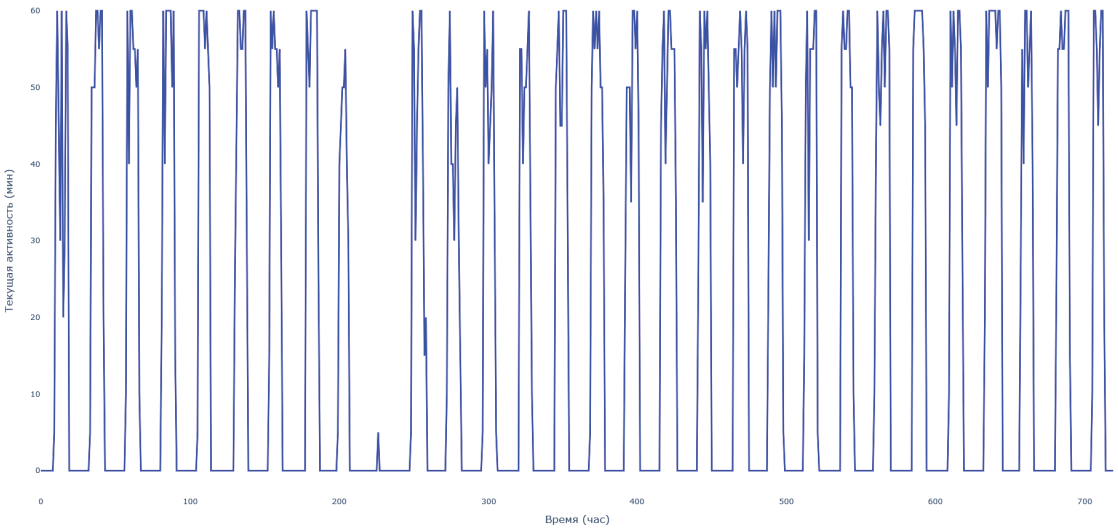

Рис.2. Базовый временной ряд

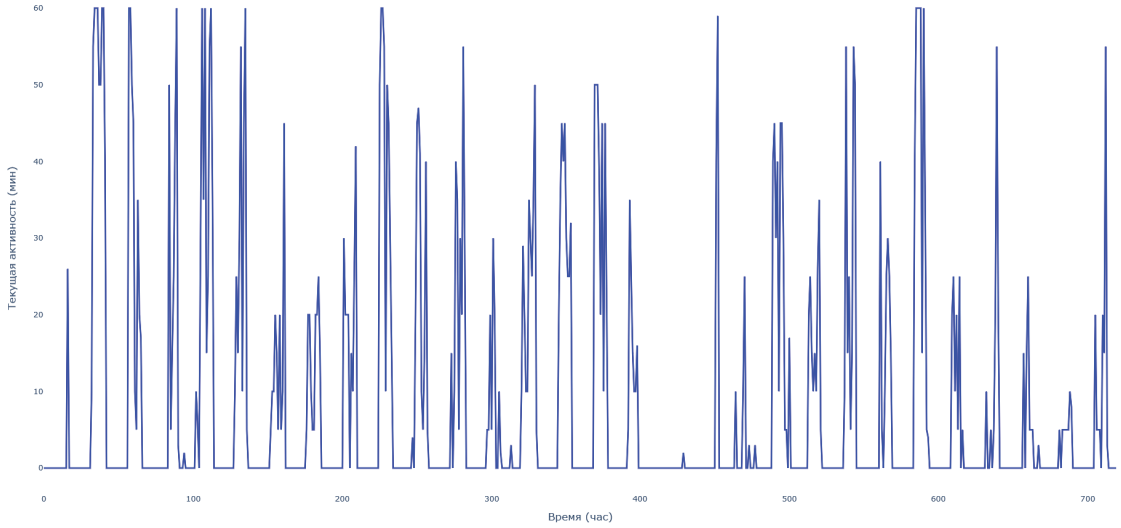

Рис.3. Текущий временной ряд

работы, время работы в программах, время работы с веб-ресурсами различных категорий (корпоративные ресурсы, социальные сети, поиск работы и т.п.), количество входящих и исходящих писем, количество низкоприоритетных и высокоприоритетных событий безопасности и др. Для примера были выбраны два временных ряда, характеризующих время активной работы одного пользователя с измеряемым промежутком, соответствующим времени рабочей недели, агрегированным по часу (рис.2,3).

На рис. 2, 3 приведены данные системы оперативного мониторинга для базового профиля активности пользователя[23] и текущие наблюдения за его активностью в течение одного месяца (по оси абсцисс – единицей измерения являются «часы», промежуток измерения – месяц; по оси ординат – единицей измерения являются минуты активного времени работы в течение часа).

Согласно методике, изложенной в работе [7], указанные временные ряды были преобразованы в трехмерные облака данных, представленные на рис. 4, 5.

Для данных множеств были построены симплициальные фильтрации Вьеториса–Рипса [5] и соот-

23 Пользователь корпоративной сети выбирался случайным образом.

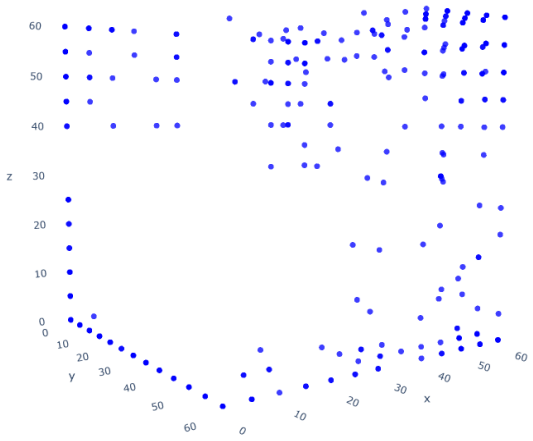

Рис.4. Облако точек базового временного ряда

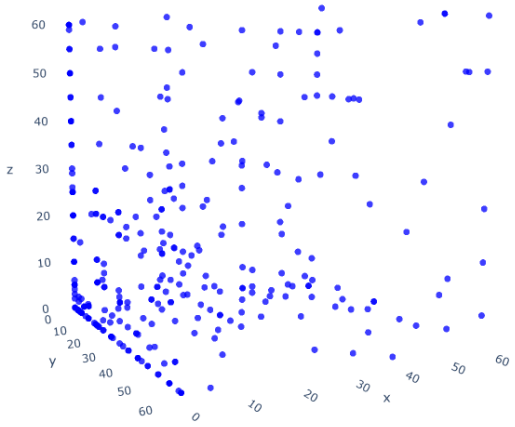

Рис.5. Облако точек текущего временного ряда

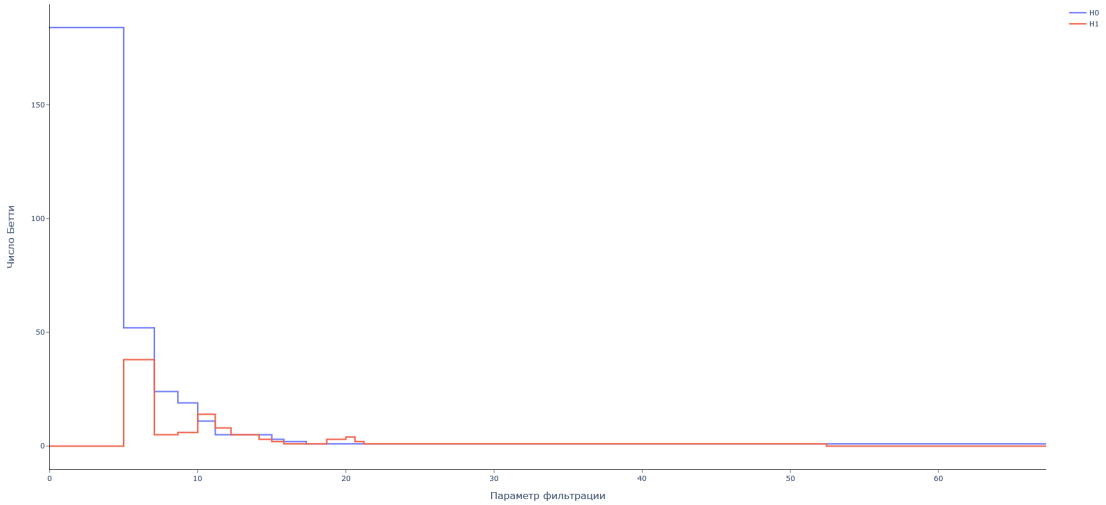

Рис.6. Кривые Бетти базового временного ряда

ветствующие кривые Бетти для нулевой ($p=0$) и первой ($p=1$) групп гомологий (группы p – мерных симплициальных гомологий, для которых $p>1$ в фильтрации отсутствуют). Они приведены на рис. 6, 7.

Классификатор для обнаружения поведенческих паттернов строился на основе модифицированных функций желательности.

Для этого, в соответствии с выражением (5), были вычислены метрики между текущей и базовой кривыми Бетти (функции $\beta_{p,f}(\cdot)$, $\beta^0_{p,f}(\cdot)$ соответственно, $p=0,1$) которые затем были пересчитаны в частные показатели желательности.

Пересчет метрик в частные показатели желательности осуществлялся в соответствии с выражением:

$$\psi_p = \exp\left\{-\exp\left\{-\left[9\left(\frac{c-\mathrm{d}_p\left(\beta_{p,f}(\cdot),\beta^0_{p,f}(\cdot)\right)}{c}\right)^{1.927}\right]\right\}\right\}, p=0,1. \quad (6)$$

где $\beta^0_{p,f}(\cdot)$ — кривая Бетти, соответствующая базовому профилю.

Параметр «c» определяется экспериментальным путем и соответствует максимальному значению метрик, полученных при сравнении средних кривых Бетти $\overline{\beta}_{p,f}(\cdot)$, $p=0,1$ и кривых Бетти, построенных на обучающей выборке. Обобщенный показатель жела-

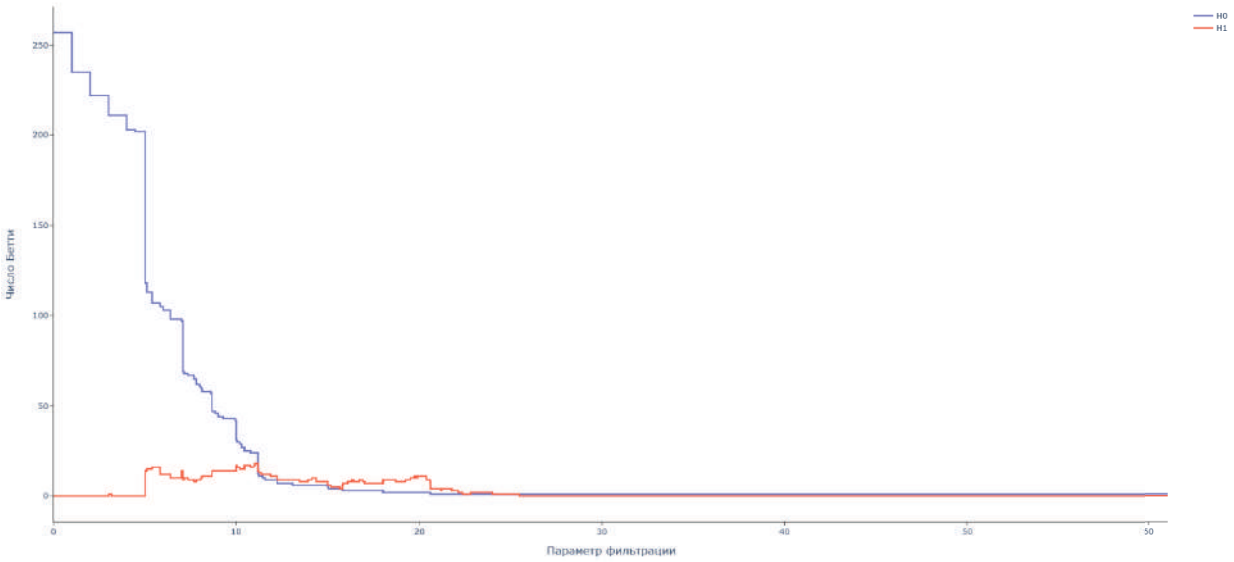

Рис.7. Кривые Бетти текущего временного ряда

тельности — D_{Σ} рассчитывается как среднее по Колмогорову с логарифмической функцией[24] от частных показателей желательности ψ_p, $p = 0,1$:

$$D_{\Sigma} = \sqrt{\psi_0 \psi_1}. \quad (7)$$

Таблица 1

Топологические дескрипторы и оценка активности пользователя информационных ресурсов

Параметр	d_0	d_1	ψ_0	ψ_1	$D_{£}$
Значение	0,8107	0,9339	0,9586	0,0028	0,05196

Решающее правило формулируется следующим образом: активность пользователя информационных ресурсов в корпоративной сети в масштабе шкалы желательности классифицируется как отклонение от типового шаблона поведения, если $D_{\Sigma} \le 0.37$.

Для приведенных выше наборов данных результаты расчетов сведены в таблицу 1. Из ее анализа следует, что в рассматриваемом случае однозначно фиксируется отклонение активности пользователя информационных ресурсов от его типового шаблона, что, при прочих равных условиях, потенциально может представлять собой инцидент безопасности.

4. Заключение

Предложенные в работе топологические дескрипторы позволяют построить с приемлемой вычислительной сложностью простые и достаточно эффективные процедуры выявления поведенческих паттернов в задачах обеспечения кибербезопасности. Перспективными направлениями дальнейшей работы представляется апробация предложенного подхода для решения задач выявления и классификации паттернов различного типа в киберфизических системах различного назначения.

Научный руководитель: Заборовский Владимир Сергеевич доктор технических наук, профессор, Санкт-Петербургский политехнический университет Петра Великого. E-mail:vladimir.zaborovsky@spbstu.ru

24 Адлер Ю.А., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. 2-е изд., перераб. и доп. —М.: Наука, 1976 —280 с.

Литература

1. Каширина И.Л., Демченко М.В. Исследование и сравнительный анализ методов оптимизации, используемых при обучении нейронных систем//Вестник ВГУ, Серия: системный анализ и информационные технологии. 2018. № 4. С.123-132. DOI: https://doi.org/10.17308/sait.2018.4/1262.
2. Sadowski G., Litan A., Bussa T., Phillips T. Market Guide for User and Entity Behavior Analytics. Published: 23 April 2018. ID: G00349450. Gartner. 2018.
3. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. – М.: МЦМНО. 2018. – 484 с.
4. Chazal F., Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv:1710.04019v2 [math.ST]. 2021. URL: https://arxiv.org/pdf/1710.04019.pdf. DOI: https://doi.org/10.48550/arXiv.1710.04019.
5. Murugan J., D. Robertson. An introduction to Topological Data Analysis for Physicists: From LGM to FRBs. arXiv:1904.11044v1. 2019. URL: https://arxiv.org/pdf/1904.11044.pdf. DOI: https://doi.org/10.48550/arXiv.1904.11044.
6. Орлов Г.М., Игнатьева О.А., Васин А.Г., Низомутдинов Б.А. Современные методы обработки и анализа данных. – СПб.: Университет ИТМО. 2021. – 147 с.
7. Нашивочников Н.В., Пустарнаков В.Ф. Топологические методы анализа в системах поведенческой аналитики. // Вопросы кибербезопасности. 2021 №2 (42). С. 26-36. DOI: 10.21681/2311-3456-2021-2-26-36.
8. Wasserman L., Topological data analysis // Annual Review of Statistics and Its Application, 2018. v.5, pp. 501–532. DOI: 10.1146/annurev-statistics-031017-100045.
9. Гринес В.З., Гуревич Е.Я., Жужома Е.В., Починка О.В. Классификация систем Морса–Смейла и топологическая структура несущих многообразий // Успехи математических наук. – 2019. т. 74, вып. 1(445), с. 41–116. DOI: https://doi.org/10.4213/rm9855.
10. Шарафутдинов В.А. Введение в дифференциальную топологию и риманову геометрию: учеб. пособие / Новосиб. гос. ун-т. – Новосибирск: ИПЦ НГУ. 2018. – 282 с.
11. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. — М.: Наука, 1989. — 624 с.
12. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407 [cs.LG]. 2019. URL: https://arxiv.org/pdf/1901.03407.pdf.
13. Pun C.S., Xia K., Lee S.X. Persistent-Homology-based Machine Learning and its Applications – A Survey. arXiv preprint arXiv:1811.00252, 2018.
14. Sheffar D. Introductory Topological Data Analysis. arXiv:2004.04108v1 [math.HO]. 2020. URL: https://arxiv.org/pdf/2004.04108.pdf. DOI: https://doi.org/10.48550/arXiv.2004.04108.
15. Carlsson G., Topological methods for data modelling // Nature Reviews Physics 2. 2020. № 697.
16. Huntsman S., Palladino J., Robinson M. Topology in cyber research. arXiv:2008.03299 [math.AT]. 2020. URL: https://arxiv.org/pdf/2008.03299.pdf.
17. Trevor J. Bihl, Robert J. Gutierrez, Kenneth W. Bauer, Bradley C. Boehmke, Cade Saie. Topological Data Analysis for Enhancing Embedded Analytics for Enterprise Cyber Log Analysis and Forensics // Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020. P. 1937-1946. DOI: 10.24251/HICSS.2020.238.
18. Tauzin G., Lupo U., Tunstall L., P´erez . B.J., Caorsi M., Medina-Mardones A.M., Dassatti A., Hess K. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. arXiv:2004.02551v2 [cs.LG]. 2021. URL: https://arxiv.org/pdf/2004.02551.pdf. DOI: https://doi.org/10.48550/arXiv.2004.02551.
19. Atienza N., Gonzalez-D´ıaz R., Soriano-Trigueros M. On the stability of persistent entropy and new summary functions for TDA. arXiv:1803.08304v7 [cs.IT]. 2020. URL: https://arxiv.org/pdf/1803.08304.pdf. DOI: https://doi.org/10.48550/arXiv.1803.08304.

IDENTIFICATION OF DEVIATIONS IN BEHAVIORAL PATTERNS OF USERS OF CORPORATE INFORMATION RESOURCES USING TOPOLOGICAL FEATURES

Nashivochnikov N.[25]

Purpose of the article: *The paper examines the application of topological data analysis to identify and classify patterns of user behavior of corporate information resources in cybersecurity tasks.*

Method: *the study is based on mathematical concepts of the theory of persistence homologies (simplex complexes, persistence diagrams, filtering functions, topological descriptors), theoretical models of metric spaces.*

Result: *A formalized definition of topological features is given, properties and relation of Betti curves with*

25 Nikolay Nashivochnikov, CISSP, Deputy General Director - Technical Director, Gazinformservice LLC, St. Petersburg, Russia. E-mail: cto@gaz-is.ru

persistence diagrams and other known topological descriptors used in topological data analysis are disclosed. The results of practical testing on the time-series data presented in the monitoring system of corporate network users' work with corporate information resources, confirm that the use of positively defined step functions to construct Betti curves provides an acceptable computational complexity of the procedures required to identify and classify user behavioral patterns. The approach proposed based on topological descriptors and modified desirability functions Harrington functions allows to reliably capture user activity deviation from a typical behavioral pattern, potentially constituting a security incident.

Scientific novelty: *is the application of topological data analysis using Betti curves to identify and classify user behavior patterns of corporate information resources.*

Keywords: *cybersecurity, UEBA, topological data analysis, Betti curves, time series.*

References

1. Kashirina I.L., Demchenko M.V. Issledovanie i sravnitel'nyj analiz metodov optimizacii, ispol'zuemyh pri obuchenii nejronnyh sistem//Vestnik VGU, Serija: sistemnyj analiz i informacionnye tehnologii. 2018. № 4. S.123-132. DOI: https://doi.org/10.17308/sait.2018.4/1262.
2. Sadowski G., Litan A., Bussa T., Phillips T. Market Guide for User and Entity Behavior Analytics. Published: 23 April 2018. ID: G00349450. Gartner. 2018.
3. V'jugin V.V. Matematicheskie osnovy mashinnogo obuchenija i prognozirovanija. – M.: MCMNO. 2018. – 484 s.
4. Chazal F., Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv:1710.04019v2 [math.ST]. 2021. URL: https://arxiv.org/pdf/1710.04019.pdf. DOI: https://doi.org/10.48550/arXiv.1710.04019.
5. Murugan J., D. Robertson. An introduction to Topological Data Analysis for Physicists: From LGM to FRBs. arXiv:1904.11044v1. 2019. URL: https://arxiv.org/pdf/1904.11044.pdf. DOI: https://doi.org/10.48550/arXiv.1904.11044.
6. Orlov G.M., Ignat'eva O.A., Vasin A.G., Nizomutdinov B.A. Sovremennye metody obrabotki i analiza dannyh. – SPb.: Universitet ITMO. 2021. – 147 s.
7. Nashivochnikov N.V., Pustarnakov V.F. Topologicheskie metody analiza v sistemah povedencheskoj analitiki. // Voprosy kiberbezopasnosti. 2021 №2 (42). S. 26 36. DOI: 10.21681/2311-3456-2021-2-26-36.
8. Wasserman L., Topological data analysis // Annual Review of Statistics and Its Application, 2018. v.5, pp. 501–532. DOI: 10.1146/annurev-statistics-031017-100045.
9. Grines V.Z., Gurevich E.Ja., Zhuzhoma E.V., Pochinka O.V. Klassifikacija sistem Morsa–Smejla i topologicheskaja struktura nesushhih mnogoobrazij // Uspehi matematicheskih nauk. – 2019. t. 74, vyp. 1(445), s. 41–116. DOI: https://doi.org/10.4213/rm9855.
10. Sharafutdinov V.A. Vvedenie v differencial'nuju topologiju i rimanovu geometriju: ucheb. posobie / Novosib. gos. un-t. – Novosibirsk: IPC NGU. 2018. – 282 s.
11. Kolmogorov A.N., Fomin S.V. Jelementy teorii funkcij i funkcional'nogo analiza. — M.: Nauka, 1989. — 624 s.
12. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407 [cs.LG]. 2019. URL: https://arxiv.org/pdf/1901.03407.pdf.
13. Pun C.S., Xia K., Lee S.X. Persistent-Homology-based Machine Learning and its Applications – A Survey. arXiv preprint arXiv:1811.00252, 2018.
14. Sheffar D. Introductory Topological Data Analysis. arXiv:2004.04108v1 [math.HO]. 2020. URL: https://arxiv.org/pdf/2004.04108.pdf. DOI: https://doi.org/10.48550/arXiv.2004.04108.
15. Carlsson G., Topological methods for data modelling // Nature Reviews Physics 2. 2020. № 697.
16. Huntsman S., Palladino J., Robinson M. Topology in cyber research. arXiv:2008.03299 [math.AT]. 2020. URL: https://arxiv.org/pdf/2008.03299.pdf.
17. Trevor J. Bihl, Robert J. Gutierrez, Kenneth W. Bauer, Bradley C. Boehmke, Cade Saie. Topological Data Analysis for Enhancing Embedded Analytics for Enterprise Cyber Log Analysis and Forensics // Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020. P. 1937-1946. DOI: 10.24251/HICSS.2020.238.
18. Tauzin G., Lupo U., Tunstall L., P´erez . B.J., Caorsi M., Medina-Mardones A.M., Dassatti A., Hess K. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. arXiv:2004.02551v2 [cs.LG]. 2021. URL: https://arxiv.org/pdf/2004.02551.pdf. DOI: https://doi.org/10.48550/arXiv.2004.02551.
19. Atienza N., Gonzalez-D´ıaz R., Soriano-Trigueros M. On the stability of persistent entropy and new summary functions for TDA. arXiv:1803.08304v7 [cs.IT]. 2020. URL: https://arxiv.org/pdf/1803.08304.pdf. DOI: https://doi.org/10.48550/arXiv.1803.08304.