DOI: 10.15514/ISPRAS-2024-36(5)-8

Дилемма защитника: совместимы ли методы защиты от разных атак на модели машинного обучения?

[1,2] *Г.В. Сазонов, ORCID: 0009-0003-0905-534X <sazonovg@ispras.ru>*
[1, 3, 4] *К.С. Лукьянов, ORCID: 0009-0009-5235-2175 <lukyanov.k@ispras.ru>*
[2] *И.Н. Мелешин, ORCID: 0009-0009-1541-3418 <igor.meleshin@graphics.cs.msu.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*
[2] *Московский государственный университет имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д. 1.*
[3] *Московский физико-технический институт (НИУ), Россия 117303, Москва, ул. Керченская, д.1 А, корп. 1*
[4] *Исследовательский центр доверенного искусственного интеллекта ИСП РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

Аннотация. В условиях растущего применения моделей искусственного интеллекта (ИИ) всё больше внимания уделяется вопросам доверия и безопасности систем, использующих ИИ от разных типов угроз (атаки уклонения, отравления, вывод о членстве и т.д.). В этой работе мы сосредотачиваемся на задаче классификации вершин графов, выделяя ее как одну из самых сложных. Эта работа является первой, насколько нам известно, в которой исследуется взаимосвязь методов защиты моделей ИИ от разных типов угроз на графовых данных. Наши эксперименты проводятся на наборах данных: цитирования и графов покупок. Мы показываем, что в общем случае нельзя просто использовать комбинации методов защит от разных типов угроз и, что это может иметь серьезные негативные последствия вплоть до полной потери эффективности модели. Мы также приводим теоретическое доказательство противоречия класса методов защиты от атак отравления на графах и состязательного обучения.

Ключевые слова: Атаки на модели искусственного интеллекта; защищенность; классификация вершин графов; доверенный искусственный интеллект.

Для цитирования: Сазонов Г.В., Лукьянов К.С., Мелешин И.Н. Дилемма защитника: совместимы ли методы защиты от разных атак на модели машинного обучения? Труды ИСП РАН, том 36, вып. 5, 2024 г., стр. 109–126. DOI: 10.15514/ISPRAS–2024–36(5)–8.

Благодарности: Работа выполнена при поддержке грантом для исследовательских центров в области искусственного интеллекта, представленным Аналитическим центром в соответствии с договором о предоставлении субсидии (идентификатор договора 000000D730321P5Q0002) и договором с Институтом системного программирования им. В. П. Иванникова от 02 ноября 2021 г. № 70-2021-00142.

The Defender's Dilemma: Are Defense Methods Against Different Attacks on Machine Learning Models Compatible?

[1, 2] *G.V. Sazonov ORCID: 0009-0003-0905-534X <sazonovg@ispras.ru>*
[1, 3, 4] *K.S. Lukyanov ORCID: 0009-0009-5235-2175 <lukyanov.k@ispras.ru>*
[2] *I.N. Meleshin ORCID: 0009-0009-1541-3418 <igor.meleshin@graphics.cs.msu.ru>*
[1] *Ivannikov Institute of System Programming of the Russian Academy of Sciences, 25, A. Solzhenitsyn str., Moscow, 109004, Russia.*
[2] *Lomonosov Moscow State University, Leninskie Gory, 1, Moscow, 119991, Russia.*
[3] *Moscow Institute of Physics and Technology (National Research University), building 1, 1 A, Kerchenskaya str., Moscow, 117303, Russia.*
[4] *Research Center for Trusted Artificial Intelligence ISP RAS, 25, A. Solzhenitsyn str., Moscow, 109004, Russia.*

Abstract. With the increasing use of artificial intelligence (AI) models, more attention is being paid to issues of trust and security in AI systems against various types of threats (evasion attacks, poisoning, membership inference, etc.). In this work, we focus on the task of graph node classification, highlighting it as one of the most complex. To the best of our knowledge, this is the first study exploring the relationship between defense methods for AI models against different types of threats on graph data. Our experiments are conducted on citation and purchase graph datasets. We demonstrate that, in general, it is not advisable to simply combine defense methods for different types of threats, as this can lead to severe negative consequences, including a complete loss of model effectiveness. Furthermore, we provide theoretical proof of the contradiction between defense methods against poisoning attacks on graphs and adversarial training

Keywords: AI model attacks; security; graph node classification; trusted AI.

For citation: G.V. Sazonov, Lukyanov K.S., Meleshin I.N. The Defender's Dilemma: Are Defense Methods Against Different Attacks on Machine Learning Models Compatible? *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 5, 2024, pp. 109-126 (in Russian). DOI: 10.15514/ISPRAS-2024-36(5)-8.

Acknowledgements: The work was supported by a grant for research centers in the field of artificial intelligence, presented by the Analytical Center in accordance with the subsidy agreement (000000D730321P5Q0002 agreement identifier) and the agreement with the Ivannikov Institute for System Programming dated November 02, 2021, No. 70-2021-00142.

1. Введение

Устойчивость глубоких нейронных сетей к входным возмущениям является важнейшим свойством для их интеграции в различные отрасли, требующие безопасности, такие как беспилотные автомобили, медицинская диагностика и финансы. Хотя нейронные сети должны выдавать схожие результаты для схожих входных данных, давно известно, что они уязвимы для состязательных возмущений [1] – небольших, вычисленных преобразований входных данных, которые не изменяют семантику входного объекта, но заставляют модель выдавать предопределенное решение. Наиболее популярный класс атак – это атаки уклонения [2-4] направленные на обман моделей во время эксплуатации. Большинство методов изучения состязательной устойчивости нейронных сетей направлены на создание состязательных возмущений, которые указывают на то, что в целом прогнозы нейронной сети ненадежны.

Также, известны атаки отравления [5-6] направленные на модификацию наборов данных, обучение на котором может приводить к общему снижению качества модели или появлению у нее так называемых бэкдоров, триггеров, появление которого в данных провоцирует модель сделать определенный неверный прогноз выгодный нарушителю. Против атак отравления в литературе предложены свои методы защиты [7-8], который позволяет выполнить особую

предобработку данных перед обучением, чтобы избавиться от подозрительных данных и/или модифицировать данные.

Несмотря на эффективность методов защиты от различных атак определенного класса, против которого они были разработаны, на данный момент нет четких рекомендаций как эффективно защитится от нескольких атак одновременно. В качестве наиболее простого и очевидного решения можно предложить просто комбинировать лучшие методы защиты от классов угроз, против которых хочется иметь защиту. Однако, в этом исследовании мы показываем, что такой подход может быть малоэффективным и приводить к серьезным негативным последствиям. Причем такой подход помимо того, что не позволяет защититься от атак различной природы, но и разрушает защиту от атак, которым модель с одним типом защит успешно противостояла, а в худшем случае модель вообще становится непригодной для использования даже без каких-либо угроз и атак.

Наш вклад можно представить следующим образом:

1. Насколько нам известно, мы первые кто исследует совместимость методов защит от разных типов угроз на графовых данных;
2. Мы показываем, что в общем случае комбинирование методов защит от разных типов атак приводит к противоречию методов защиты, ухудшая модель и не позволяя эффективно защититься ни от одной из угроз;
3. Мы приводим теоретическое доказательство противоречия класса методов защиты от атак отравления на графах и состязательного обучения;
4. Мы экспериментально подтверждаем все выкладки на двух графовых доменах.

Структура статьи организована следующим образом. Во втором разделе представлен обзор литературы, в котором рассматриваются основные подходы и результаты, достигнутые в данной области исследования. Третий раздел посвящен постановке задач, где формулируются основные определения и цели исследования, которые оно призвано решить. В четвертом разделе описывается методология, применяемая для достижения поставленных целей. Пятый раздел содержит описание проведенных экспериментов и анализ полученных результатов. В шестом разделе рассматриваются ограничения предложенных подходов и их потенциальное влияние на результаты. Наконец, седьмой раздел содержит заключение, в котором подводятся итоги работы, а также намечаются возможные направления для будущих исследований.

2. Обзор литературы

В этом разделе приводится краткий обзор существующих методов интерпретации, атак черного ящика и защит от состязательных атак применительно к моделям, работающим с графовыми данными.

Существуют разные подходы к интерпретации данных. Наиболее распространенным в литературе является апостериорный подход к интерпретации, когда модель сначала обучается, а уже только после этого применяются методы интерпретации. Методы апостериорной интерпретации для моделей, работающих с графовыми данными, можно разделить на три основные группы: методы, основанные на внимании, распространении значимости и градиентные методы. Каждый подход по-своему объясняет предсказания графовых моделей, уделяя особое внимание топологии и структуре графов.

Все атаки можно разделить на разные классы угроз, которые различаются моментом атаки в большом цикле разработки и эксплуатации моделей машинного обучения, а также, по задачи атаки. Так атаки уклонения [9-11] направлены на нарушение работы модели путём манипуляции входными данными на этапе эксплуатации модели. Атаки отравления модифицируют данные до обучения существуют атаки, нацеленные на общее снижение эффективности модели после обучения [5-6] и есть атаки отравления, нацеленные на

внедрения триггеров [12-13]. Также, есть атаки, направленные на нарушение приватности, например, класс атак вывода о членстве [14-15], цель которых понять какие данные использовались для обучения модели. Есть атаки похищения модели [16], цель которых извлечь примерные веса закрытой модели посредством запросов или обучить свою модель затратив значительно меньшее вычислительные мощности, чем требуется для обучения похищаемой модели. Есть и другие классы атак [17-18]. Однако, в этой работе мы сосредоточились на исследовании атак отравления, понижающих общую эффективность после обучения и атаках уклонения.

Целевые и нецелевые атаки на узлы и подграфы – эти методы нацелены на изменение конкретных узлов и рёбер, чтобы повлиять на предсказания. Nettack [10] – пример целевой атаки, которая изменяет рёбра и признаки отдельных узлов для нарушения предсказаний. FGSM (Fast Gradient Sign Method) [9], является примером целевой атаки, применяющей градиентные оценки для создания искажений на уровне вершин графов. Переносимые атаки – атаки, направленные на создание искажений, которые будут эффективны для нескольких моделей одновременно. Атаки чёрного ящика – атаки, где злоумышленник не имеет доступа к параметрам модели и полагается на изменения в предсказаниях, чтобы провести атаку. RL-S2V [11] – это метод, использующий обучение с подкреплением для атаки на графовые модели без знания их внутренних параметров, полагаясь только на доступ к предсказаниям модели.

Эти подходы к построению различных типов атак уклонения показывают, как изменения в графовой структуре могут нарушить предсказания, подчёркивая необходимость в разработке надёжных защитных механизмов, способных противостоять таким атакам.

Атаки CLGA [6] и Mettack [5] являются примерами атак отравления, направленных на графовые данные, с различными стратегиями для достижения максимального воздействия на модели графового обучения. CLGA – метод ориентирован на использование контрастивных потерь для внесения искажений в графовые данные на этапе обучения. Этот подход основан на том, чтобы злоумышленник изменял рёбра графа, минимизируя качество представлений, получаемых в ходе обучения методом без учителя. Это достигается с помощью обратного распространения градиентов через функцию контрастивных потерь, что делает искажения более целенаправленными и эффективными, даже в отсутствии меток. Mettack – атака отравления, основанная на метаобучении, предложенная в работе [5]. Основная идея состоит в том, чтобы оптимизировать структуру графа или признаки его узлов с учетом предстоящего обучения модели. Mettack моделирует процесс обучения целевой модели как вложенную оптимизационную задачу, где внешняя оптимизация направлена на внесение искажений, ухудшающих итоговые предсказания. Этот метод поддерживает как целевые, так и нецелевые атаки и может быть применен к различным архитектурам графовых нейронных сетей, что делает его переносимым.

Эти подходы к построению различных типов атак отравления показывают, как изменения в графовой структуре могут снижать ее эффективность при обучении после применения этих типов атак на сырые данные.

Причем, методы, которые модифицируют данные после обучения модели с целью обмануть модель являются атаками уклонения, а методы атаки, применяемые к исходным данным до обучения модели, являются атаками отравления.

Методы защиты от атак на графы можно классифицировать по их подходам к устойчивости: защита на уровне данных, включающая фильтрацию и корректировку графа, архитектурные методы, повышающие устойчивость моделей к искажениям, и регуляризация, предотвращающая зависимость от отдельных узлов или рёбер. Основные стратегии включают фильтрацию, добавление шума и устойчивые к атакам архитектуры. Основное внимание в этой работе будет уделено методом, позволяющим противостоять состязательным атакам.

Фильтрация и корректировка данных – метод, который на уровне данных предотвращает добавление искажений, фильтруя подозрительные ребра и узлы. Jaccard Similarity Filtering [8] выполняет фильтрацию рёбер на основе схожести признаков узлов, препятствуя внедрению вредоносных рёбер и относится к защите от атак отравления. GNN Guard [7] использует механизмы для обнаружения подозрительных рёбер и снижает их влияние на предсказания, что защищает от атак и также относится к методам защиты от атак отравления. Архитектурные изменения – изменение модели для повышения устойчивости к атакам. Robust GCN [19] добавляет регуляризацию и шум на уровне узлов, снижая чувствительность модели к мелким искажениям. Регуляризация и шум – методы, предотвращающие переобучение на отдельных элементах графа, что делает модель менее восприимчивой к атакам. Также к регуляризации можно отнести Adversarial training [20] и Gradient Regularization [21], которые меняют процесс обучения добавляя данные применением атаки во время обучения и ограничивая градиенты, защищая модель от атак уклонения.

Эти защитные подходы позволяют значительно снизить уязвимость графовых моделей от соответствующих типов угроз, сохраняя высокую точность предсказаний и надежность модели при использовании графов.

Соответственно методы защиты, который выполняет модификацию данных до обучения является методами защиты от атак отравления. Методы защит, которые модифицируют поведение/архитектуру модели и/или меняют процедуру обучения, являются методами защиты от атак уклонения.

На основании обзора литературы видно, что существует множество методов атак и защит. В работе [22] было проведено исследования о противоречии методов защит, для моделей классификации изображений. Однако, на данный момент в литературе не представлено как одновременно защищаться от атак разного типа при этом сохранив высокую эффективность, а также не известно характерна ли проблема несовместимости методов защит от разных типов данных для моделей, работающих с другими типами данных.

3. Постановка задач

В этом разделе формально вводится постановка задач, вводятся обозначения, используемые в статье, и формулируются вопросы исследования. Введем формальные определения атак уклонения, отравления, защищенных моделей от соответствующих атак и частично защищенной модели.

3.1 Атаки на модель классификации

Атака уклонения (англ. *evasion attack*) для моделей классификации – это процесс, при котором злоумышленник стремится изменить входные данные $x \in R\hat{}d$ таким образом, чтобы модифицированные данные $x' = x + \delta$ обошли классификацию модели $f: R\hat{}d \to Y$, где Y – множество классов.

Целью атаки уклонения является нахождение возмущения $\delta \in R\hat{}d$, удовлетворяющего следующим условиям:

1. $arg⟦max⟧\_(y \in Y)\ f(x') \neq arg⟦max⟧\_(y \in Y)\ f(x)$, то есть предсказанный класс изменяется после применения возмущения.
2. $\| \delta \| \_p \leq \epsilon$, где $\|\cdot\| \_p$ – L_p норма (например, с $p = 2$ или $p = \infty$), а ϵ – заданное ограничение на величину возмущения, чтобы сохранялась правдоподобность измененного входа.

Таким образом, задача атаки уклонения может быть формализована как задача оптимизации: $arg⟦max⟧\_(y \in Y)\ f(x') \neq arg⟦max⟧\_(y \in Y)\ f(x)$, при условии $\| \delta \| \_p \leq \epsilon$

Атаки уклонения делятся на два основных типа:

1. Белый ящик (white-box): Злоумышленник обладает полным доступом к модели f, включая её параметры и градиенты.
2. Чёрный ящик (black-box): Злоумышленник имеет ограниченный доступ к модели f, например, может наблюдать только выходы модели в ответ на определенные входы.

Стоит отметить, что в случае графовых данных атака может модифицировать вместо матрицы признаков матрицу ребер графа $G(V, E)$, где V – количество вершин, E – количество ребер или обе матрицы одновременно.

Атака отравления (англ. *poisoning attack*) – это процесс, при котором злоумышленник модифицирует обучающие данные модели машинного обучения с целью ухудшить её производительность или добиться определённого поведения на новых данных.

Пусть $D\_train = \{(x\_i, y\_i)\}\_(i = 1)\^n$ – исходный набор обучающих данных, где $x\_i \in R\^d$ – входные данные, а $y\_i \in Y$ – метки классов. Атака отравления предполагает создание нового набора данных $D\_poisoned = \{(x\_i', y\_i')\}\_(i = 1)\^n$, такого что:

Обученная на $D\_poisoned$ модель $f\^{}$ демонстрирует ухудшение качества на проверочных данных D_test:

$$L(f\^{}, D\_test) > L(f, D\_test),$$

где L – функция ошибки, а f – модель, обученная на чистых данных D_train.

1. Либо модель $f\^{}$ демонстрирует целенаправленное поведение, выгодное злоумышленнику, например, ошибочную классификацию целевого входа x_target:

$$arg\ (max)\┬(y \in Y)\ f\^{}\ (x\_target) = y\_adversarial,$$

2. где $y\_adversarial$ – целевой класс, заданный злоумышленником.

Задача атаки отравления может быть формализована как задача оптимизации: найти $L(f\^{}, D\_test) > \tau$, или $arg\ (max)\┬(y \in Y)\ f\^{}\ (x\_target) = y\_adversarial$, где τ – порог допустимой ошибки на проверочных данных.

Атаки отравления классифицируются на два основных типа:

- **Целенаправленные (targeted):** Злоумышленник модифицирует данные с целью добиться определенного поведения модели на конкретных входах.
- **Общие (untargeted):** Злоумышленник стремится ухудшить обобщающую способность модели на всех проверочных данных.

3.2 Защищенная модель

Модель $f: R\^d \to Y$ называется защищенной от атак уклонения, если она сохраняет свою устойчивость к малым возмущениям входных данных. Формально, для любого входа $x \in R\^d$ и любого допустимого возмущения $\delta \in R\^d$, удовлетворяющего $\| \delta \| \_p \leq \epsilon$, модель удовлетворяет следующему условию:

$$arg\ (max)\┬(y \in Y)\ f(x + \delta) = arg\ (max)\┬(y \in Y)\ f(x),$$

где $\epsilon > 0$ – заданное ограничение на норму возмущения, а $\|\cdot\| \_p$ – норма p (например, $p = 2$ или $p = \infty$).

Таким образом, защищённая модель f остается инвариантной к возмущениям, которые находятся в пределах заданного ограничения ϵ, обеспечивая корректную классификацию даже в условиях возможного противодействия злоумышленника.

Модель $f: R\^d \to Y$ называется защищенной от атак отравления, если она сохраняет свою способность к корректной классификации проверочных данных $D\_test = \{(x\_i, y\_i)\}\_(i = 1)\^m$, даже если обучающие данные D_train содержат модифицированные элементы, созданные злоумышленником.

Формально, пусть $D\_poisoned = D\_train \cup D\_adv$, где $D\_adv = \{(x\_i', y\_i')\}\_(i = 1)\^k$ – добавленные или измененные злоумышленником данные. Модель f считается защищенной, если:

$$L(f, D\_test) \leq L\_clean + \delta,$$

где L – функция ошибки, L_clean – ошибка модели, обученной на чистых данных, а $\delta \geq 0$ – допустимое отклонение, ограничивающее влияние атакующего.

Кроме того, для целенаправленных атак защищённая модель должна обеспечивать корректную классификацию целевых данных x_target:

$$arg\ (max)\_(y \in Y)\ f(x\_target) = y\_true,$$

где y_true – истинная метка класса.

Стоит отметить, что полной защиты удается добиться в редких случаях, тогда можно говорить, что модель является **частично защищенной**, если:

$$(i)\ \ Q(f, X) - Q(f', X) \leq \alpha,$$
$$(ii)\ \ Q(f', X') - Q(f, X') \geq \beta,$$

где $Q(\cdot,\cdot)$ – функция качества модели на входных данных, α – максимальное допустимое снижение качества на чистых данных, а β – минимально допустимое увеличение качества на атакованных данных относительно незащищенной модели. f – исходная модель, а f' – её защищённая версия. Обозначим $X \in R\^d$ как чистые входные данные, а X' – атакованные данные. Эти условия обеспечивают частичную устойчивость модели к атакам, сохраняя приемлемый уровень производительности.

Вопросы исследования

- Можно ли взять произвольные качественные методы защиты в своих категориях (от атак уклонения или от атак отравления) и получить защищенную модель от обоих типов угроз?
- Какие негативные последствия могут возникать при комбинировании методов защит?

В следующем разделе мы отвечаем на эти вопросы и подкрепляем это соответствующими экспериментами.

4. Методология

В данном разделе представлена методика исследования, направленная на достижение поставленных целей и решение заявленных задач.

4.1 Математическое противоречие методов защиты

В этом подразделе математически показано, что добавление защиты от атак отравления математически противоречит защите состязательного обучения от атаки уклонения на структуру графа.

Пусть задан граф $G(V, E)$, где V – вершины в графе, E - ребра в графе , определена атака уклонения Ev посредством удаления и добавления ребер с параметром $\epsilon; 0 < \epsilon \ll 1$, то есть атака Ev может удалить $k\_1 < E \vee ; G(V, E)$ ребер и добавить $k\_2 < E\_full/E \vee ; G\_full\ ((V, E\_full))/G\ (V, E)$ (где G_full клика построенная на вершинах V, а E_full количество ребер в клике) ребер так, чтобы $0 < k\_1 + k\_2 = k < \epsilon * E$. Состязательная защита AT использует для генерации состязательных примеров ту же атаку Ev с тем же параметром ϵ и определен метод защиты от атак отравления PD с параметром удаления $t; 0 < t < 1$. Далее для простоты количество ребер будет обозначаться как E, а не |E|.

Теорема. При одновременном использовании метода защит от атак отравления PD и состязательного обучения AT от атак уклонения типа Ev верны два утверждения: 1) с увеличением параметра t метода защиты от атак отравления PD уменьшается количество возможных состязательных примеров, которые могут быть найдены во время выполнения состязательного обучения AT; 2) с увеличением параметра t метода защиты от атак отравления PD уменьшается корреляция между распределением количества ребер возмущенных графов во время применения состязательного обучения AT и распределением количества ребер возмущенных графов полученных методом Ev во время выполнения атаки уклонения .

Доказательство 1 утверждения. Параметр t говорит о том, что метод защиты от атак отравления модифицирует граф G посредством удаления из него до $t * E$ ребер. Обозначим количество удаленных ребер как $0 < m * E \le t * E; 0 < m \le t < 1$. В результате получается граф $G'(V; (1-m)E)$. Тогда общее количество атак N_attack (модификаций графа отличных от изначального графа G), можно вычислить по следующей формуле:

$$N_{attack} = \sum_{k_1=0}^{k} \left(\binom{E}{k_1} * \sum_{k_2=0}^{k-k_1} \binom{\frac{V*(V-1)}{2} - E}{k_2} \right),$$

где функция $\binom{n}{k} = C_k^n$ и соответствует биномиальному коэффициенту, определяя количества сочетаний из n по k.

В каждом слагаемом первой суммы фиксируется значение k_1 пробегая значения от 0 до k. При каждом фиксированном k_1 надо выбрать k_1 ребер из E для удаления, что определяет биномиальный коэффициент $\binom{E}{k_1}$ и это число умножается на все допустимые варианты добавить k_2 ребер выбранных из дополнения графа G до клики G_full, так как в клике на V вершин $(V * (V-1))/2$ ребер, из которых надо вычесть существующие ребра E, то общее количество таких вариантов при фиксированном k_2 равно $\binom{\frac{V*(V-1)}{2} - E}{k_2}$. k_2 в свою очередь может принимать любое значение от 0 до $k - k\_1$ при заранее определенном значении k_1.

При добавлении метода защиты от атак отравления PD бюджет атаки Ev используемой состязательной защитой AT станет меньше, обозначим бюджет атаки Ev при условии применения метода защиты от атаки отравления PD как $k' = \epsilon * (1-m) * E = \epsilon * E' < k = \epsilon * E$. Теперь заметим, что формула для вычисления N_attack зависит от трех параметров: $V; E; \epsilon$, в случае атаки уклонения, обозначим эту величину за $N\_attack\text{^}evasion$. И от четырех: $V; E; \epsilon; m$, в случае одновременного использования двух методов защит: $AT; PD$, обозначим эту величину как $N\_attack\text{^}(full-defense)$. Тогда можно заметить, что $N\_attack\text{^}evasion > N\_attack\text{^}(full-defense)$. Распишем явно обе формулы:

$$N_{attack}^{evasion}(V; E; \epsilon) = \sum_{k_1=0}^{\epsilon*E} \left(\binom{E}{k_1} * \sum_{k_2=0}^{\epsilon*E-k_1} \binom{\frac{V*(V-1)}{2} - (1-\epsilon)*E}{k_2} \right)$$

$$N_{attack}^{full-defense}(V; E; \epsilon; m) = \sum_{k_1=0}^{\epsilon*(1-m)*E} \left(\binom{E}{k_1} * \sum_{k_2=0}^{\epsilon*(1-m)*E-k_1} \binom{\frac{V*(V-1)}{2} - (1-\epsilon)*E}{k_2} \right)$$

Функция N_attack является убывающей функцией при уменьшении E. Так как $0 < m \leq t < 1$, то $N\_attack\hat{}evasion > N\_attack\hat{}(full - defense)$, что требовалось доказать в первой части теоремы.

Доказательство 2 утверждения.

Пусть вероятность добавить ребро во время генерации состязательного примера методом Ev равна p, вероятность удалить ребро q и вероятность не изменить граф 1-p-q. Тогда случайная величина θ равная изменению количеству ребер в состязательном примере G_attack, полученного в ходе применения метода Ev как атаки уклонения, по отношению к количеству ребер в изначальном графе G имеет среднее и дисперсию:

$$E(\theta) = ((-1) * p + 1 * q + 0 * (1 - p - q))\epsilon E = E\epsilon(q - p)$$
$$E(\theta\hat{}2) = (1 * p + 1 * q + 0 * (1 - p - q))\epsilon E = E\epsilon(q + p)$$
$$V(\theta) = E\epsilon(q + p) + E\epsilon$$

При применении метода защиты PD количество ребер уменьшится на $0 < mE < tE < E$. Тогда перед применением метода Ev в состязательном обучении количество ребер в графе станет равно $E(1 - m)$. Тогда случайная величина ξ равная изменению количеству ребер в состязательном примере G_at, полученного в ходе применения метода Ev в процессе выполнения состязательного обучения AT, по отношению к количеству ребер в изначальном графе после применения метода защиты от атак отравления PD имеет среднее и дисперсию:

$$E(\xi) = ((-1) * p + 1 * q + 0 * (1 - p - q))\epsilon E(1 - m) = E\epsilon(1 - m)(q - p)$$
$$V(\xi) = E\epsilon(1 - m)(q + p) + E\epsilon(1 - m)$$

Случайная величина ξ складывается из $\epsilon(1 - m)E$ случайных реализаций выбора, добавления или пропуску действия, а θ из ϵE. Между ними есть пересечение, равное $\epsilon(1 - mE)$, т.е. ξ полностью входит в θ. Остальная часть θ содержит ϵmE уникальных членов. Ковариация Cov определяется как сумма ковариаций всех пересекающихся членов, так как ковариация между непересекающимися членами равна нулю. Посчитаем ковариацию и корреляцию случайных величин ξ и θ

$$Cov(\xi; \theta) = V(\xi)$$

$$r(\xi; \theta) = \frac{Cov(\xi;\theta)}{\sqrt{V(\xi) * V(\theta)}} = \frac{E\epsilon(1-m)\left(p+q-p^{2}+2pq-q^{2}\right)}{\sqrt{E\epsilon(1-m)\left(p+q-p^{2}+2pq-q^{2}\right)E\epsilon\left(p+q-p^{2}+2pq-q^{2}\right)}} = \frac{1-m}{\sqrt{1-m}} = \sqrt{1 - m}$$

Корреляция $r(\xi; \theta)$ является убывающей функцией с ростом m. При стремлении m к 1 получается полное отсутствие корреляции, а при m стремящимся к 0 получается корреляция близкая к 1, что и требовалось доказать.

Следствие из теоремы 1. Различие структур между возмущенными графами, получаемыми в результате применения состязательного обучения и возмущениями, генерируемыми во время атаки уклонения может привести к тому, что состязательное обучение обеспечит защиту от атак уклонения, но не в той области, в которой необходимо. Это следует из основного принципа состязательного обучения, что модель дообучается на примерах из того же распределения, что и примеры, которые может найти атака. Однако, как было показано выше, добавление защиты от атак отравления построенной на принципе удаления ребер приводит к смещению распределения и к уменьшению количества примеров, которые могут быть найдены.

4.2 Методика комбинирования методов защиты от разных типов угроз

В этой работе были рассмотрены методы атак отравления: CLGA [6] и Mettack [5]; метод атаки уклонения: FGSM [9]; метод защиты от атак отравления: Jaccard Similarity Filtering [8];

методы защиты от атак уклонения: Adversarial training [20] и Gradient Regularization [21]. Все методы, которые были предложены для других типов данных, были адаптированы для работы с графовыми моделями решающие задачу классификации вершин. Однако, на нашей реализации атаки отравления Mettack не удалось воспроизвести удовлетворительные результаты из оригинальной статьи, поэтому из экспериментов он был исключен.

Рассмотрим в какой момент цикла разработки моделей машинного обучения появляются угрозы атак отравления и уклонения, а также, в какой момент необходимо добавлять методы защиты от соответствующих типов угроз. Атака происходит перед обучением модели, после этого должен идти метод защиты от атак отравления, который выполняет функцию очистки/модификаций данных от потенциальных некачественных данных. Атака уклонения, как упоминалось ранее, происходит на этапе эксплуатации модели, что в рамках лабораторных исследований равносильно атаки на тестовый набор данных. Чтобы защититься от атак уклонения необходимо модифицировать процесс обучения. Например, состязательное обучение (Adversarial training) расширяет обучающую выборку самостоятельно сгенерированными атаками, которые генерируются по ходу обучения модели.

Чтобы оценить, как методы защиты влияют друг на друга, сначала измерялось качество модели на исходных данных, затем при наличии только атаки одного типа. Далее измерялось качество при наличии защиты на исходных данных. После чего проводилось обучение с атакой и использованием защиты от атак выбранного типа. И последнее обучение было при наличии обоих типов атак (атак отравления и атак уклонения) и обоих методов защиты соответственно. Эксперименты с атакой и защитой только от другого типа атак не проводились, так как если данные были модифицированы атакой отравления, то защита от атак уклонения не сможет их исправить из-за того, что эти методы предполагают, что данные для обучения качественные. Аналогично наличие защиты от атаки отравления не поможет защититься от атак уклонения, так как атаки уклонения используют градиенты обученной модели, чтобы вычислить шумовую маску для обмана модели на тесте, а так как защита от атак отравления модифицирует данные, то это изменит конечную модель, но не сделает ее устойчивой к шумовой маске, вычисленной на основании градиентов.

5. Эксперименты

В этом разделе описаны эксперименты и все необходимое для их воспроизведения.

5.1 Математическое противоречие методов защиты

5.1.1 Датасеты и обучение моделей

В экспериментах использовались датасеты: Cora – относящийся к домену цитирования [23] и представленных в библиотеке torch-geometric в группе датасетов Planetoid и Photo – домена графов покупок [24], также представленные в torch-geometric в группе датасетов Amazon.

Статистика наборов данных:

- Cora: 2708 вершин, 10556 ребер, 7 классов, 1433 признаков
- Photo: 7650 вершин, 238162 ребер, 7 классов, 745 признаков

В качестве модели обучалась двухслойная модель GCN (GCN-2l) и двухслойная модель GIN (GIN-2l).

```
GCN-2l(
  Sequential(
    (0): GCNConv(input_size, 16)
```

```
            (1): ReLU(inplace)
            (2): GCNConv(16, output_size)
            (3): LogSoftmax(inplace)
        )
    )



    GIN-2l(
        Sequential(
            (0): GINConv(
                Sequential(
                    (0): Linear(input_size, 16)
                    (1): BatchNorm1d(16, eps=1e-05)
                    (2): ReLU(inplace)
                    (3): Linear(16, 16)
                    (4): BatchNorm1d(16, eps=1e-05)
                    (5): ReLU(inplace)
                )
            )
            (1): ReLU(inplace)
            (2): GINConv(
                Sequential(
                        (0): Linear(16, 16) (1):
                        (1): BatchNorm1d(16, eps=1e-05)
                        (2): ReLU(inplace)
                        (3): Linear(16, output_size)
                    )
                )
            (3): LogSoftmax(inplace)
        )
    )
```

Для обучения использовался оптимизатор Adam с параметрами по умолчанию и функция ошибки NLLLoss из библиотеки PyTorch, разделение на батчи не используется. Количество эпох обучения равно 200 для достижения высокой точности классификации на всех наборах данных. Данные были разделены в соотношении 80 к 20 для обучения и тестирования соответственно.

5.1.2 Параметры проведения экспериментов

В экспериментах использовались: метод атаки отравления: CLGA [6]; метод атаки уклонения: FGSM [9]; метод защиты от атак отравления: Jaccard Similarity Filtering [8]; методы защиты от атак уклонения: Adversarial training [20] и Gradient Regularization [21]. Гиперпараметры соответствующих методов представлены в табл. 1.

5.1.3 Протокол оценки

В качестве основной метрики оценки моделей использовалась метрика точности (Accuracy). Для оценки эффективности атак и защит использовалась разница точности моделей при наличии или отсутствии соответствующего метода. Также, в случае неоднозначности выводов при использовании атак по одной метрики использовалась метрика ASR (Average Success Rate; средняя вероятность успеха). Для методов защиты смотрелось как на изменение метрики точности как на исходных данных, так и на атакованных для комплексной оценки

влияния метода защиты на модель.

Табл. 1. Гиперпараметры методов атак и защит.
Table 1. Hyperparameters of attack and defense methods.

Метод	Гиперпараметры
CLGA	learning_rate = 0.01 num_hidden = 256 num_proj_hidden = 32 activation = *prelu* drop_edge_rate_1 = 0.3 drop_edge_rate_2 = 0.4 tau = 0.4 num_epochs = 3000 weight_decay = $1e-5$ drop_scheme = *degree*
FGSM	$\epsilon = 0.005$
Jaccard	threshold = 0.4
Adversarial training	attack_name = *FGSM* $\epsilon = 0.01$
Gradient Regularization	regularization_strength = 50

5.2 Результаты экспериментов

При наличии атаки отравления результат экспериментов усредняются по 5 запускам (так как обучение одной модели с применением атаки отравления занимает до 10 часов), в остальных случаях по 15.

В табл. 2, 3, 4 и 5 представлены результаты экспериментов с использованием различных комбинаций методов атак и защит. Сокращенные названия методов: AdvTrain – защита состязательного обучения; GradReg – защита регуляризацией градиентов. В табл. 2 и 3 представлены результаты на наборе данных Cora, домен цитирования. В табл. 4 и 5 представлены результаты на наборе данных Photo, домен граф покупок. В табл. 2 и 4 представлены результаты модели GCN-2l. В табл. 3 и 5 представлены результаты модели GIN-2l.

Можно заметить, что комбинирования защит на исходных данных не вызывает никаких особых негативных последствий (верхний левый квадрат результатов каждой таблицы), а в некоторых случаях комбинирование защит даже приводит к повышению точности относительно результатов при отдельном использовании каждого из методов защиты. При этом отметим, что каждый из методов защит по отдельности приводит к повышению эффективности при противодействии от своей профильной угрозы. Также заметим, что как только появляются угрозы, то результаты моделей значимо падают при использовании комбинации защит относительно случаев, когда от угрозы защищаются профильным методом защиты (левый нижний и правый верхний квадраты результатов в каждой таблице). Также, видно, что при комбинации угроз качество значимо падает и в 3-х случаях из 4-х становится еще ниже, чем при наличии только одной угрозы (правый нижний квадраты результатов в каждой таблице).

В табл. 2 показана точность модели GCN-2l на наборе данных Cora при различных комбинациях методов атак и защит. Пропуски, обозначенные символом "\textemdash" означают, что эксперимент с соответствующей конфигурацией не проводился. В первой строке указан метод атаки уклонения, его отсутствие подписано как "No attack". В первом столбце указан метод атаки отравление, его отсутствие подписано как "No attack". Во второй

строке указан метод защиты от атак уклонения или его отсутствие, которое помечено как "No defense". Во втором столбце указан метод защиты от атак отравления или его отсутствие, которое помечено как "No defense". В левом верхнем углу указан набор данных и модель.

Табл. 2. Точность модели GCN -2l на наборе данных Cora при различных комбинациях методов атак и защит.
Table 2. Accuracy of the GCN -2l model on the Cora dataset under various attack and defense methods combinations.

Cora	GCN-2l	No attack			FGSM		
		No defense	AdvTrain	GradReg	No defense	AdvTrain	GradReg
No attack	No defense	90 ± 0	88 ± 2	91 ± 1	57 ± 3	78 ± 3	70 ± 4
	Jaccard	77 ± 5	81 ± 5	80 ± 5	46 ± 4	69 ± 5	63 ± 2
CLGA	No defense	71 ± 2	—	—	—	—	—
	Jaccard	76 ± 6	67 ± 3	61 ± 2	—	57 ± 4	42 ± 1

В табл. 3 показана точность модели GIN-2l на наборе данных Cora при различных комбинациях методов атак и защит. Пропуски, обозначенные символом "\textemdash" означают, что эксперимент с соответствующей конфигурацией не проводился. В первой строке указан метод атаки уклонения, его отсутствие подписано как "No attack". В первом столбце указан метод атаки отравление, его отсутствие подписано как "No attack". Во второй строке указан метод защиты от атак уклонения или его отсутствие, которое помечено как "No defense". Во втором столбце указан метод защиты от атак отравления или его отсутствие, которое помечено как "No defense". В левом верхнем углу указан набор данных и модель.

Табл. 3. Точность модели GIN-2l на наборе данных Cora при различных комбинациях методов атак и защит.
Table 3. Accuracy of the GIN-2l model on the Cora dataset under various attack and defense methods combinations.

Cora	GIN-2l	No attack			FGSM		
		No defense	AdvTrain	GradReg	No defense	AdvTrain	GradReg
No attack	No defense	90 ± 3	86 ± 2	85 ± 3	43 ± 3	76 ± 1	59 ± 4
	Jaccard	65 ± 4	80 ± 6	75 ± 4	29 ± 3	66 ± 3	53 ± 4
CLGA	No defense	66 ± 5	—	—	—	—	—
	Jaccard	64 ± 3	65 ± 2	39 ± 7	—	52 ± 2	37 ± 3

В табл. 4 показана точность модели GCN-2l на наборе данных Photo при различных комбинациях методов атак и защит. Пропуски, обозначенные символом "\textemdash" означают, что эксперимент с соответствующей конфигурацией не проводился. В первой строке указан метод атаки уклонения, его отсутствие подписано как "No attack". В первом столбце указан метод атаки отравление, его отсутствие подписано как "No attack". Во второй строке указан метод защиты от атак уклонения или его отсутствие, которое помечено как "No defense". Во втором столбце указан метод защиты от атак отравления или его отсутствие, которое помечено как "No defense". В левом верхнем углу указан набор данных и модель.

В табл. 5 показана точность модели GIN-2l на наборе данных Photo при различных комбинациях методов атак и защит. Пропуски, обозначенные символом "\textemdash" означают, что эксперимент с соответствующей конфигурацией не проводился. В первой строке указан метод атаки уклонения, его отсутствие подписано как "No attack". В первом столбце указан метод атаки отравление, его отсутствие подписано как "No attack". Во второй строке указан метод защиты от атак уклонения или его отсутствие, которое помечено как "No defense". Во втором столбце указан метод защиты от атак отравления или его отсутствие, которое помечено как "No defense". В левом верхнем углу указан набор данных и модель.

Табл. 4. Точность модели GCN-2l на наборе данных Photo при различных комбинациях методов атак и защит.
Table 4. Accuracy of the GCN-2l model on the Photo dataset under various attack and defense methods combinations.

Photo	GCN-2l	No attack			FGSM		
		No defense	AdvTrain	GradReg	No defense	AdvTrain	GradReg
No attack	No defense	94 ± 1	91 ± 1	92 ± 2	90 ± 2	92 ± 2	92 ± 1
	Jaccard	92 ± 2	92 ± 3	92 ± 2	89 ± 1	88 ± 2	87 ± 1
CLGA	No defense	88 ± 2	—	—	—	—	—
	Jaccard	90 ± 1	88 ± 1	87 ± 1	—	85 ± 1	84 ± 1

Табл. 5. Точность модели GIN-2l на наборе данных Photo при различных комбинациях методов атак и защит.
Table 5. Accuracy of the GIN-2l model on the Photo dataset under various attack and defense methods combinations.

Photo	GIN-2l	No attack			FGSM		
		No defense	AdvTrain	GradReg	No defense	AdvTrain	GradReg
No attack	No defense	83 ± 2	85 ± 2	78 ± 7	62 ± 13	81 ± 3	73 ± 3
	Jaccard	65 ± 5	66 ± 7	60 ± 4	56 ± 4	60 ± 10	53 ± 7
CLGA	No defense	58 ± 6	—	—	—	—	—
	Jaccard	64 ± 3	70 ± 8	58 ± 5	—	78 ± 3	65 ± 2

5.3 Обсуждение результатов

Из результатов экспериментов можно сделать выводы, что комбинирование произвольных методов защиты может не просто не помогать защититься от двух и более угроз, а приводить к ухудшению качества и разрушению защиты даже от одной атаки, против которой без дополнительного метода защиты модель успешно справлялась.

Также можно отметить специфические результаты на датасете Photo. В случае модели GCN-2l падение качества было не слишком большим. Это можно объяснить тем, что гиперпараметры всех атак и защит подбирались на обучающей выборке датасета Cora и оставались неизменными для набора данных Photo. Для получения более существенных изменений качества моделей необходимо увеличить возмущения для FGSM атак и увеличить количество итераций для атаки отравления CLGA (так как количество итераций определяет, сколько возмущений в данные может внести этот метод, корректнее вносить изменения в процентном отношении от общего числа элементов графа, а не в абсолютном). А в случае модели GIN-2l можно заметить, что при нескольких угрозах качество растет по сравнению с тем, когда угроза только одна. Это можно объяснить тем, что в силу структурных особенностей графов покупок удаление ребер не позволяет найти уникальные паттерны, так как свертка GIN работает на принципе поиска изоморфных подграфов и в отличии от случая с датасетом Cora атака, скорее всего, приводит к невозможности выделения уникальных изоморфных графов.

6. Ограничения и обсуждение

Стоит отметить, что доказанная в этой работе теорема имеет ограниченное применение и не учитывает возможность атаки не только структуры графа, но и атаки на признаки. Также, класс защиты от отравления ограничен теми, которые только удаляют подозрительные данные. Однако, теории, которая бы однозначно могла сказать, как непротиворечиво надо комбинировать методы защит, на данный момент нет, и предложенная теорема – это первый шаг к созданию рекомендаций, подкрепленных не только экспериментами, но и

математической теорией.

Хочется также отметить, что возможным способом решить проблему смещения распределения при комбинировании методов защит может быть увеличение допустимых возмущений во время состязательного обучения. С практической точки зрения это означает, что если мы считаем атаку незаметной, если ее бюджет возмущений был 5% от всего графа, то можно, например, разрешить возмущения в 10% во время состязательного обучения, таким образом распределение генерируемых возмущенных графов с учетом смещения при комбинировании будет включать в себя и распределение с допустимым возмущением в 5%, но которое не было смещено.

Отдельно отметим, что основной вывод этой работы, что комбинирование методов защиты, который по отдельности были эффективные не гарантирует защиту при комбинации. При этом этот вывод не говорит, что невозможно защитится от нескольких угроз, а говорит, что следует исследовать эту проблематику в будущих работах и аккуратно подходить к построению систем, требующих наличие защищенности от нескольких различных типов угроз.

7. Заключение и будущая работа

В этой статье было впервые показано, что комбинирование произвольных методов защит от разных типов угроз не позволяет защититься от них на графовых данных. Причем наивное комбинирование не просто не позволяет получить модель, защищенную от обеих угроз одновременно, но и может приводить к серьезным последствиям разрушая защиту от угрозы, от которой ранее была защищена. Мы показываем верность этих утверждений для нескольких различных архитектур и нескольких доменов графов (графах цитирования и графах покупок). А также, приведено теоретическое доказательство как класса методов защиты от атак отравления на графах мешает обеспечить защиту посредством состязательного обучения от атак уклонения.

В качестве направлений будущей работы можно выделить исследования большего количества методов и разработку практических рекомендаций эффективных комбинаций, принципы обеспечения защит которых не разрушают защиты друг друга, что может быть оформлено в большой бенчмарк. Более того, можно расширить бенчмарки исследованиями противодействия трем и более типам угроз. Также, можно проверить наличие этой проблемы на других типах данных: текстовые данные, временные ряды и видео. Еще одним направлением может быть создание методов защит, которые изначально были бы способны противостоять сразу нескольким типам угроз, при этом не имеющих внутренних противоречий принципов обеспечения защиты от разных типов угроз. А также, стоит развивать теоретическую базу анализирующую возможность комбинировать методы защит, на основании которой можно будет формировать рекомендации позволяющие обеспечивать одновременную защиту от атак разного типа.

Список литературы / References

[1]. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*.

[2]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

[3]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[4]. Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2574–2582.

[5]. Zügner, D., & Günnemann, S. (2019). Adversarial Attacks on Graph Neural Networks via Meta Learning. In *International Conference on Learning Representations, Workshop Track*. Available at: https://arxiv.org/abs/1902.08412.
[6]. Zhang, S., Chen, H., Sun, X., Li, Y., & Xu, G. (2022). Unsupervised graph poisoning attack via contrastive loss back-propagation. In *Proceedings of the ACM Web Conference 2022*, 1322–1330.
[7]. Zhang, X., & Zitnik, M. (2020). Gnnguard: Defending graph neural networks against adversarial attacks. In *Advances in neural information processing systems*, 33, 9263–9275.
[8]. Wu, H., Wang, C., Tyshetskiy, Y., Docherty, A., Lu, K., & Zhu, L. (2019). Adversarial examples on graph data: Deep insights into attack and defense. *arXiv preprint arXiv:1903.01610*.
[9]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
[10]. Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2847–2856.
[11]. Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., & Song, L. (2018). Adversarial attack on graph structured data. In *International conference on machine learning*, 1115–1124.
[12]. Zhang, Z., Jia, J., Wang, B., & Gong, N. Z. (2021). Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, 15–26.
[13]. Zheng, H., Xiong, H., Chen, J., Ma, H., & Huang, G. (2022). Motif-Backdoor: Rethinking the Backdoor Attack on Graph Neural Networks via Motifs. *arXiv preprint arXiv:2210.13710*.
[14]. Shaikhelislamov, D., Lukyanov, K., Severin, N., Drobyshevskiy, M., Makarov, I., & Turdakov, D. (2024). A study of graph neural networks for link prediction on vulnerability to membership attacks. *Journal of Mathematical Sciences*, 1–11.
[15]. Conti, M., Li, J., Picek, S., & Xu, J. (2022). Label-Only Membership Inference Attack against Node-Level Graph Neural Networks. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, 1–12.
[16]. Yuan, X., Ding, L., Zhang, L., Li, X., & Wu, D. O. (2022). Es attack: Model stealing against deep neural networks without data hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5), 1258–1270.
[17]. Wang, S., & Gong, Y. (2022). Adversarial example detection based on saliency map features. *Applied Intelligence*, 52(6), 6262–6275.
[18]. Ma, J., Deng, J., & Mei, Q. (2022). Adversarial attack on graph neural networks as an influence maximization problem. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 675–685.
[19]. Zhu, D., Zhang, Z., Cui, P., & Zhu, W. (2019). Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 1399–1407.
[20]. Feng, F., He, X., Tang, J., & Chua, T.-S. (2019). Graph adversarial training: Dynamically regularizing based on graph structure. *IEEE Transactions on Knowledge and Data Engineering*, 33(6), 2493–2504.
[21]. Finlay, C., & Oberman, A. M. (2019). Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*.
[22]. Szyller, S., & Asokan, N. (2023). Conflicting interactions among protection mechanisms for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), 15179–15187.
[23]. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, 29(3), 93–93.
[24]. McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Информация об авторах / Information about authors

Георгий Владимирович САЗОНОВ – сотрудник отдела информационных систем института системного программирования им. В.П. Иванникова Российской академии наук; студент магистратуры МГУ.

Georgii Vladimirovich SAZONOV – an employee of the Information Systems Department of the

Ivannikov Institute for System Programming of the Russian Academy of Sciences; master's student at Moscow State University.

Кирилл Сергеевич ЛУКЬЯНОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН; аспирант МФТИ.

Kirill Sergeevich LUKYANOV – Researcher at the Center for Trusted Artificial Intelligence of the Ivannikov Institute for System Programming of the Russian Academy of Sciences; postgraduate student at Moscow Institute of Physics and Technology.

Игорь Николаевич МЕЛЕШИН – сотрудник лаборатории компьютерной графики и мультимедиа; студент бакалавриата МГУ.

Igor Nikolaevich MELESHIN – employee of the Laboratory of Computer Graphics and Multimedia Moscow State University; undergraduate student of Moscow State University.