**РЕЦЕНЗИЯ**

върху дисертационен труд за придобиване на образователната и научна степен „Доктор“
Научна област: 5. Технически науки
Професионално направление: 5.13. „Общо инженерство“
Докторска програма: „Компютърни технологии в инженерната дейност“

Автор на дисертационния труд: ас. маг. инж. Иван Любомиров Дрънков

Тема на дисертационния труд: „АНАЛИЗ НА ИЗТЕКЛИ КРИТИЧНИ ДАННИ С ЦЕЛ РАЗПОЗНАВАНЕ НА ПОТЕНЦИАЛНИ ЗАПЛАХИ ЗА ИНФОРМАЦИОННАТА СИГУРНОСТ“

*Научен ръководител:* доц. д-р Веселин Иванов Христов

*Рецензент:* доц. д-р Георги Запрянов, ТУ-София, ФКСТ, кат. „Компютърни системи“

**1. Актуалност на разработвания в дисертационния труд проблем в научно и научно-приложно отношение**

Темата на дисертационния труд е актуална и практически значима. Разглежда се проблемът за последващото използване на вече компрометирана информация като източник за оценяване на риска и ранно разпознаване на потенциални заплахи. С увеличаването на обемите от лични, корпоративни и технически данни нараства и възможността изтекли потребителски идентификатори, пароли, токени, конфигурационни файлове, метаданни и други чувствителни ресурси да бъдат използвани за фишинг, злоупотреба с идентификационни данни, кражба на самоличност, неоторизиран достъп и целенасочени атаки срещу организации.

Авторът правилно поставя акцент върху обстоятелството, че традиционните механизми за защита са насочени предимно към предотвратяване и откриване на инцидента, докато последващият анализ на вече изтекли данни често изисква ръчна работа с OSINT инструменти, специализирани скриптове и разнородни информационни източници. В този контекст разработването на интегрирана среда за събиране, верификация, унифициране, анализ, известяване и генериране на доклади е основателна научно-приложна задача.

Основната цел е разработване и прилагане на методологичен подход за автоматизирана обработка, унифициране и анализ на големи обеми хетерогенни изтекли данни и създаване на концептуален модел на Активна система и анализатор на изтекли данни (АСАИД). За реализиране на целта са формулирани пет задачи, обхващащи особеностите на хетерогенните масиви, стандартизирането на данните, архитектурния модел, алгоритмите за анализ и тестовото им изследване. Целта и задачите са логически свързани и съответстват на тематиката на докторската програма.

Дисертационният труд е в обем 131 страници и включва увод, четири глави, заключение, приноси и библиография от 56 източника. В текста са посочени 78 фигури и 9 таблици. Значителният практически обем на изследването се определя и от обработката на масиви, достигащи приблизително 5 милиарда записа и около 2 TB информация.

**2. Степен на познаване състоянието на проблема и творческа интерпретация на литературния материал**

Дисертантът демонстрира добро познаване на проблематиката, свързана с изтичането на данни, OSINT, защитата на личните данни, обработката на големи информационни масиви, клиент-сървър архитектурите, разпределените бази данни и приложимите аналитични процедури. Библиографията съдържа 56 източника – научни публикации, монографии, нормативни и технически материали, документация на използвани технологии и официални ресурси на разглежданите системи.

В първа глава са анализирани съществуващи услуги за работа с компрометирани данни и е предложено разграничение между пасивни системи за известяване, системи за известяване и активни системи за анализ. Подробно са разгледани HIBP, Dehashed, Snusbase, Firefox Dark Web Monitoring, LastPass Dark Web Monitoring и F-Secure Identity Theft Checker, като авторът посочва, че в по-широкото проучване са обхванати над 30 подобни платформи. Съществен положителен момент е стремежът да се открие функционален дефицит в съществуващите решения и върху него да се обоснове необходимостта от АСАИД.

Литературният и технологичният обзор е достатъчен за поставяне на задачата, но на места е по-скоро описателен. Част от твърденията са основани на продуктова документация и уеб ресурси, а критичното сравнение между различни научни и технологични подходи би могло да бъде по-систематично. При динамично променящите се онлайн услуги е желателно сравнението да бъде обвързано с ясно фиксиран момент на проверка и еднакви критерии за оценка. Тези бележки не променят извода, че авторът познава добре изследваната област и е способен да интерпретира материала спрямо поставения проблем.

**3. Съответствие на избраната методика на изследване с поставената цел и задачи**

Избраният подход следва последователност, подходяща за поставената цел: намиране и събиране на данни; верификация и валидиране; почистване и дедубликация; унифициране и стандартизиране; категоризиране по тип и критичност; съхранение; аналитична обработка; извеждане на потенциални заплахи; известяване и генериране на доклад. Разграничаването между валидиране на формата и структурата и верификация на достоверността на данните е методологично правилно.

Съществена част от методиката е разработеният софтуерен инструмент PyFormat, чрез който се автоматизират предварителната обработка, дедубликацията и преобразуването на хетерогенни данни към унифициран CSV формат. Архитектурната част е основана на модифицирана трислойна клиент-сървър структура, допълнена с клъстерна организация, MongoDB sharding, помощни бази, аналитични модули и модул за известяване. Този подход осигурява логическа модулност и възможност за бъдещо хоризонтално мащабиране.

Работата позволява да се разграничат концептуалното проектиране, прототипната реализация и експериментално валидираните части. Реализирани са функционални елементи на клиентския слой, сървърни услуги, база данни, модул за известяване, PyFormat, аналитични процедури и генериране на доклади. Същевременно пълната експериментална проверка на клъстерното ускорение не е извършена поради ограничената физическа инфраструктура. Това ограничение е

коректно посочено от автора и трябва да се отчита при оценката на мащабируемостта, но не компрометира адекватността на общата методика.

**4. Аналитична характеристика и оценка на достоверността на материала**

***Първа глава***

Първа глава има обзорен характер. Представени са основни понятия, връзката между изтекли данни и OSINT, правни аспекти, класификация на видовете данни и системите за работа с тях. Предложената класификация ПСИИД-СИД-АСАИД е използвана последователно като основа за формулиране на изследователския проблем. Изводът, че наличните услуги са ориентирани главно към проверка и известяване, а не към интегриран автоматизиран анализ, е логическата връзка към следващите глави. По-силна страна би било въвеждането на предварително формализирана матрица от критерии за оценяване на всички разглеждани платформи, тъй като техните функции се изменят във времето.

***Втора глава***

Втора глава развива концептуалния модел на АСАИД. Предлага се модулна трислойна архитектура с клиентски, сървърен и слой за данни, допълнени с клъстерен сървър, shared-nothing организация и специализирани модули за анализ и известяване. Самите използвани технологии – REST API, клъстеризация, виртуализация, контейнери и разпределени бази – са утвърдени решения; собствената стойност на разработката се състои в тяхното систематизиране и адаптиране към конкретния процес за обработка на изтекли критични данни.

В експерименталната част е описан минимален локален прототип с AMD Ryzen 7 3800X, 32 GB RAM и SSD за логическия слой и Dell R720xd с два Xeon E5-2665, 64 GB ECC RAM и дисков масив за слоя за данни. Реализирани са и три виртуални Ubuntu Server 24 LTS възела. Авторът изрично отбелязва, че споделените физически ресурси на виртуалните машини не позволяват обективно измерване на ускорението от клъстеризацията. Измерени са времена за заявки върху приблизително 1 TB реални данни, но стойностите след MongoDB sharding са дадени като прогнозни, а не като експериментално доказани резултати. Това е коректно ограничение, което следва да бъде ясно разграничено от валидираните резултати.

***Трета глава***

Третата глава разглежда проблемите при подготовката и управлението на данните: дублиране, нестандартни формати, кодиране, липсващи и неконсистентни записи, валидиране, верификация, категоризиране и съхранение. Формулирани са практически процедури за сравнение с референтни източници, кръстосана проверка, проверка на извадки и съпоставяне с други бази. Представена е и методика за честотен анализ на категориите критични данни.

Количествената база е значителна. В отделни части на труда обаче се срещат стойности „над 3 милиарда“ и „приблизително 5 милиарда“ записа. В анализа на честотите се използва приблизително 5-милиарден масив и са представени групи от 62%, 28% и 10%. Формулите за относителна честота са ясни, но не е достатъчно формализирано по какви прагове отделните категории се отнасят към групите с висока, средна и ниска честота. За възпроизводимостта на изследването би било полезно да се разграничат точно общо събрани, верифицирани, обработени и използвани в конкретните анализи записи.

***Четвърта глава***

Четвърта глава е основната приложна част на дисертацията. Представени са технологичният стек и прототипната реализация на АСАИД, процесите за събиране и проверка на данни, PyFormat, анализите на пароли, хешове, локации, домейни, документи и метаданни, използването на OSINT, извеждането на потенциални заплахи, модулът за известяване и автоматизираното генериране на PDF доклади. Модулът за известяване включва асинхронна обработка, обработка на изключения, криптирани канали, логване на събития и REST API интеграции, което показва по-висока степен на реализация от чисто концептуален модел.

PyFormat е практически значим резултат. Алгоритъмът включва анализ, редактиране, конфигуриране и извличане, както и дедубликация. Експериментът с CSV файл от 2 974 336 реда и 49 полета демонстрира обработка на голям масив, а посочената скорост варира от над 1 000 до около 100 000 реда в секунда. Авторът посочва, че чрез PyFormat са валидирани и приведени в CSV формат над 5 милиарда записа и че в процеса на разработката алгоритмите са ускорени над 25 пъти. Сравнението с MS Excel показва предимството спрямо конвенционален офис продукт, но за по-силна оценка на производителността би било необходимо сравнение със специализирани ETL/Big Data инструменти при еднакви условия и повторяеми измервания.

Анализът на 3 723 514 реални потребителски пароли е един от най-добре количествено подкрепените експерименти. Изследвани са дължина, честота, позиция и тип на използваните символи, като се търсят зависимости, свързани с човешкия фактор при създаване на пароли. Геолокационният анализ е реализиран с лични данни на автора, което е правилен избор от етична гледна точка. Използвани са EXIF/GPS, IP и ръчно зададени координати, Delaunay триангулация и претегляне на точки. Получена е зона с приблизителен радиус 500 метра, но самият автор отбелязва, че резултатът не може да бъде обективно верифициран. Не са достатъчно подробно обосновани и количествено валидирани теглата, присвоявани на различните източници на локационни данни.

В края на главата е представено функционално сравнение между АСАИД и съществуващи системи чрез показатели „Да/Не/Ограничено“. То показва по-широкия замислен функционален обхват на АСАИД, но не представлява количествен benchmark по отношение на латентност, пропускателна способност, точност на анализа, цена и използвани ресурси. Като цяло материалът в четвърта глава доказва наличие на реална програмна разработка и експериментална работа, макар част от архитектурните и аналитичните резултати да са прототипни или демонстрационни.

**5. Научни и научно-приложни приноси на дисертационния труд**

В дисертацията са формулирани три научно-приложни и два приложни приноса. Приемам по същество приносния характер на разработката, но считам, че формулировките могат да бъдат прецизирани така, че по-ясно да отделят новите резултати от извършените обзорни дейности.

Като най-съществени научно-приложни резултати оценявам:

- Разработената методологична последователност за преминаване от сурови хетерогенни изтекли данни към верифицирани, структурирани и аналитично използваеми масиви;
- Концептуалния модел на АСАИД, който интегрира събиране, съхранение, обработка, аналитични модули, известяване и докладване;

- Адаптирането и обединяването на различни аналитични процедури към специфичната задача за разпознаване на потенциални заплахи.

Като ясно изразени приложни приноси приемам разработването и експерименталното използване на PyFormat и частичната реализация на АСАИД с функционални модули за обработка, анализ, известяване и генериране на доклади. Формулировката „разгледани и съпоставени са основните анализи“ сама по себе си има по-скоро систематизиращ характер и не бих я определил като самостоятелен силен научно-приложен принос без обвързването ѝ с предложената методика и архитектура.

В обобщение, приносите са предимно научно-приложни и приложни и са достатъчни по характер за дисертационен труд в докторската програма „Компютърни технологии в инженерната дейност“.

**6. Оценка на личното участие на докторанта**

Представеният материал дава основания да се приеме значително лично участие на дисертанта. В текста са описани собствената разработка PyFormat, изграждането и конфигурирането на прототипни компоненти на АСАИД, събирането и обработването на големи информационни масиви, разработването на аналитични скриптове и провеждането на експерименти. При локационния анализ са използвани лични данни на автора, което допълнително показва непосредствено участие в експерименталната работа.

От трите публикации по темата една е самостоятелна, а останалите две са в съавторство. На основата на дисертационния текст и представения списък на публикациите приемам, че основните резултати и приносите са в съществена степен лично дело на кандидата.

**7. Преценка на публикациите по дисертационния труд**

Към дисертацията са представени три публикации, тематично свързани с основните направления на изследването: самостоятелна публикация от 2022 г. за използване на зависимости от изтекли данни при генериране на пароли; публикация от 2023 г. за геолокационна автентикация чрез данни от изтичания; публикация от 2026 г. за OSINT инструмент за геолокационен анализ чрез координатна триангулация, посочена като приета за печат. Всички са в Annual of the University of Mining and Geology „St. Ivan Rilski“.

Положително оценявам наличието на самостоятелна публикация и пряката тематична връзка между публикациите и дисертационния труд. В самата дисертация не е представена информация за цитирания, индексиране в международни бази данни или отделна библиометрична оценка, поради което такива твърдения не могат да бъдат направени само по предоставения труд. Съответствието с конкретните минимални национални и институционални показатели следва да се установява от пълния комплект документи по процедурата.

**8. Използване на резултатите от дисертационния труд в научната и социалната практика. Практическа приложимост**

Резултатите имат ясна практическа насоченост. Реализирани са прототипни компоненти за търсене на компрометирани идентификатори, обработка и унифициране на данни, анализ на пароли, локации, домейни, документи и метаданни, автоматично известяване и генериране на PDF доклади. Разгледана е и възможност за облачна реализация на АСАИД чрез Google Cloud,

като е направена ориентировъчна оценка на необходимата инфраструктура и месечните разходи за нея.

Практическата стойност е най-силно доказана на ниво изследователски прототип и собствен софтуерен инструмент. Не са представени документи за внедряване на АСАИД в реална организация, продължителна експлоатация или пряк икономически ефект. Поради това резултатът следва да се оценява като реализиран прототип с потенциал за практическо внедряване, а не като завършена продукционна система.

**9. Оценка на съответствието на автореферата с изискванията и адекватността на отразяване на дисертационния труд**

Авторефератът е в обем от 55 стр. и е изготвен в съответствие с академичните изисквания. След съпоставка с пълния дисертационен труд считам, че авторефератът като цяло адекватно представя структурата, поставените задачи, архитектурния модел, основните експериментални резултати, приносите и ограниченията на изследването. Основните резултати за PyFormat, анализа на пароли, локационния анализ, функционалните модули и сравнението със съществуващи системи са отразени в съответствие с дисертацията.

Налице са отделни редакционни несъответствия, включително различно посочен брой таблици – 10 в автореферата и 9 в дисертацията, както и неточности в номерацията и изписването на някои термини и наименования. Тези проблеми не засягат същността на изследването.

**10. Бележки и препоръки**

Към дисертационния труд могат да бъдат отправени следните бележки и препоръки:

- В различни части на дисертацията се посочват различни обеми на изследваните данни – „над 3 милиарда“ и „приблизително 5 милиарда“ записа. Не е направено достатъчно ясно разграничение между общо събраните, верифицираните, обработените и включените в отделните експерименти записи.
- Посочената стойност от около 95% верифицирани данни не е подкрепена с достатъчно формализиран количествен критерий. Описани са използваните подходи за кръстосана проверка и сравнение с референтни източници, но не е представена методика, позволяваща еднозначно възпроизвеждане на тази оценка.
- Предложената клъстерна архитектура на АСАИД е реализирана и изследвана основно във виртуализирана среда със споделени физически ресурси. Поради това не е експериментално доказана степента на ускорение и мащабируемост при преминаване към реална shared-nothing конфигурация, а част от оценките за производителността имат прогнозен характер.
- Експерименталните резултати за PyFormat показват работоспособност при големи информационни масиви, но сравнителната оценка на производителността е ограничена. Използването на MS Excel като основен сравнителен еталон не позволява пълноценна оценка спрямо специализирани средства за ETL и обработка на големи данни.
- При анализа на локационни данни и използването на Delaunay триангулация са въведени тегла за различните източници на координатна информация, но начинът за определяне на тези тегла и влиянието им върху крайния резултат не са количествено

валидирани. Самият автор отбелязва ограничената възможност за обективна проверка на получените резултати.

Направените бележки не намаляват стойността на дисертационния труд и имат по-скоро препоръчителен характер, като подчертават сериозния мащаб на извършената изследователска работа.

**11. Заключение**

Дисертационният труд на ас. маг. инж. Иван Любомиров Дрънков разглежда актуален и практически значим проблем в областта на информационната сигурност, анализа на данни и проектирането на мащабируеми компютърни системи. Разработката следва последователна логика – от анализ на съществуващите решения и формулиране на собствена класификация, през методология за подготовка и верификация на данните и архитектурен модел, до прототипна реализация и експериментални изследвания. Като съществени резултати оценявам концептуалния модел на АСАИД, разработения софтуерен инструмент PyFormat, систематизираната процедура за обработка на големи хетерогенни масиви, функционалните модули за анализ и известяване, анализа на значителен масив от реални потребителски пароли и автоматизираното генериране на аналитични доклади. Основното е, че разработката не се ограничава до концепция, а съдържа реално създадени и изследвани софтуерни компоненти.

Наред с това трудът има ясно очертани ограничения: клъстерното мащабиране не е валидирано върху независими физически възли; част от количествените твърдения се нуждаят от по-строга формализация; някои експерименти са демонстрационни поради законови и етични ограничения. Авторът сам отчита основна част от тези ограничения и формулира насоки за бъдещо развитие.

Представените резултати имат преимуществено научно-приложен и приложен характер и показват способността на кандидата самостоятелно да формулира инженерно-научен проблем, да проектира архитектурно решение, да разработва софтуер и да провежда експериментални изследвания с големи информационни масиви.

Считам, че дисертационният труд отговаря на изискванията на Закона за развитие на академичния състав в Република България и съответните правилници. На основание на изложеното давам положителна оценка на дисертационния труд и убедено препоръчвам на Уважаемото научно жури да присъди на ас. маг. инж. Иван Любомиров Дрънков образователната и научна степен „Доктор“.

София

24.08.2026 г. Рецензент: ….…….

/доц. д-р Георги Запрянов/

**REVIEW**

On a dissertation for the award of the educational and scientific degree "Doctor"
Field of higher education: 5. Technical Sciences
Professional field: 5.13. General Engineering
Doctoral programme: "Computer Technologies in Engineering"

Author of the dissertation: Assist. Prof. MSc Eng. Ivan Lyubomirov Drankov

Title of the dissertation: "Analysis of Leaked Critical Data for the Identification of Potential Information Security Threats"

*Scientific supervisor:* Assoc. Prof. Veselin Ivanov Hristov, PhD

*Reviewer:* Assoc. Prof. Georgi Zapryanov, PhD, Technical University of Sofia, FCST, Computer Systems Department

**1. Relevance of the problem addressed in the dissertation from a scientific and scientific-applied perspective**

The topic of the dissertation is relevant and practically significant. It addresses the subsequent use of already compromised information as a source for risk assessment and early identification of potential threats. As the volumes of personal, corporate, and technical data increase, so does the possibility that leaked user identifiers, passwords, tokens, configuration files, metadata, and other sensitive resources may be used for phishing, credential abuse, identity theft, unauthorized access, and targeted attacks against organizations.

The author appropriately emphasizes that traditional protection mechanisms are aimed primarily at preventing and detecting incidents, whereas the subsequent analysis of data that have already leaked often requires manual work with OSINT tools, specialized scripts, and heterogeneous information sources. In this context, the development of an integrated environment for collection, verification, unification, analysis, notification, and report generation is a well-founded scientific-applied task.

The main objective is to develop and apply a methodological approach for the automated processing, unification, and analysis of large volumes of heterogeneous leaked data and to create a conceptual model of an Active System and Analyzer of Leaked Data (ASAID). Five tasks have been formulated to achieve this objective, covering the characteristics of heterogeneous datasets, data standardization, the architectural model, analysis algorithms, and their experimental testing. The objective and tasks are logically related and correspond to the subject area of the doctoral programme.

The dissertation comprises 131 pages and includes an introduction, four chapters, a conclusion, contributions, and a bibliography of 56 sources. The text contains 78 figures and 9 tables. The substantial practical scope of the research is also demonstrated by the processing of datasets reaching approximately 5 billion records and about 2 TB of information.

**2. Knowledge of the state of the problem and creative interpretation of the literature**

The doctoral candidate demonstrates good knowledge of the issues related to data leakage, OSINT, personal data protection, processing of large information datasets, client-server architectures, distributed databases, and applicable analytical procedures. The bibliography contains 56 sources - scientific publications, monographs, regulatory and technical materials, documentation for the technologies used, and official resources of the systems under review.

Chapter One analyses existing services for working with compromised data and proposes a distinction between passive notification systems, notification systems, and active analysis systems. HIBP, DeHashed, Snusbase, Firefox Dark Web Monitoring, LastPass Dark Web Monitoring, and F-Secure Identity Theft Checker are examined in detail, while the author states that more than 30 similar platforms were covered in the broader study. A significant positive aspect is the effort to identify a functional gap in existing solutions and to use it to justify the need for ASAID.

The literature and technology review is sufficient for defining the research task, although in places it is predominantly descriptive. Some statements are based on product documentation and web resources, while the critical comparison between different scientific and technological approaches could be more systematic. Given the dynamic nature of online services, the comparison should be tied to a clearly specified date of assessment and uniform evaluation criteria. These remarks do not change the conclusion that the author has a good command of the research field and is capable of interpreting the material in relation to the problem addressed.

**3. Adequacy of the selected research methodology to the stated objective and tasks**

The selected approach follows a sequence appropriate to the stated objective: locating and collecting data; verification and validation; cleaning and deduplication; unification and standardization; categorization by type and criticality; storage; analytical processing; identification of potential threats; notification; and report generation. The distinction between validation of data format and structure and verification of data authenticity is methodologically sound.

A substantial part of the methodology is the developed PyFormat software tool, which automates preprocessing, deduplication, and conversion of heterogeneous data into a unified CSV format. The architectural part is based on a modified three-tier client-server structure supplemented by cluster organization, MongoDB sharding, auxiliary databases, analytical modules, and a notification module. This approach provides logical modularity and the possibility of future horizontal scaling.

The work makes it possible to distinguish between conceptual design, prototype implementation, and experimentally validated components. Functional elements of the client layer, server services, a database, a notification module, PyFormat, analytical procedures, and report generation have been implemented. At the same time, a complete experimental evaluation of cluster acceleration has not been carried out because of the limited physical infrastructure. This limitation is correctly identified by the author and should be taken into account when assessing scalability, but it does not compromise the adequacy of the overall methodology.

**4. Analytical characteristics and assessment of the reliability of the material**

***Chapter One***

Chapter One is of a review nature. It presents the main concepts, the relationship between leaked data and OSINT, legal aspects, and classifications of data types and systems for working with them. The

proposed classification of passive leaked-data notification systems, leaked-data notification systems, and ASAID is used consistently as a basis for formulating the research problem. The conclusion that existing services are mainly oriented toward checking and notification rather than integrated automated analysis provides the logical link to the subsequent chapters. A stronger aspect would have been the introduction of a predefined formalized matrix of criteria for evaluating all platforms considered, since their functionality changes over time.

***Chapter Two***

Chapter Two develops the conceptual model of ASAID. A modular three-tier architecture is proposed, comprising client, server, and data layers, supplemented by a clustered server, shared-nothing organization, and specialized analysis and notification modules. The technologies used themselves - REST APIs, clustering, virtualization, containers, and distributed databases - are established solutions; the original value of the work lies in their systematization and adaptation to the specific process of handling leaked critical data.

The experimental section describes a minimal local prototype with an AMD Ryzen 7 3800X, 32 GB RAM, and SSD for the logical layer, and a Dell R720xd with two Xeon E5-2665 processors, 64 GB ECC RAM, and a disk array for the data layer. Three virtual Ubuntu Server 24 LTS nodes are also implemented. The author explicitly notes that the shared physical resources of the virtual machines do not allow an objective measurement of the acceleration resulting from clustering. Query times are measured on approximately 1 TB of real data, but the values after MongoDB sharding are presented as projected rather than experimentally verified results. This is a correctly stated limitation that should be clearly distinguished from the validated results.

***Chapter Three***

Chapter Three addresses the problems involved in data preparation and management: duplication, non-standard formats, encoding, missing and inconsistent records, validation, verification, categorization, and storage. Practical procedures are formulated for comparison with reference sources, cross-checking, sample verification, and comparison with other databases. A methodology for frequency analysis of critical data categories is also presented.

The quantitative basis is substantial. However, different parts of the dissertation refer to "more than 3 billion" and "approximately 5 billion" records. The frequency analysis uses a dataset of approximately 5 billion records and presents groups of 62%, 28%, and 10%. The formulas for relative frequency are clear, but the thresholds by which individual categories are assigned to high-, medium-, and low-frequency groups are not sufficiently formalized. For reproducibility of the research, the total number of collected, verified, processed, and specifically analyzed records should be clearly distinguished.

***Chapter Four***

Chapter Four is the principal applied part of the dissertation. It presents the technology stack and prototype implementation of ASAID, the processes for data collection and verification, PyFormat, analyses of passwords, hashes, locations, domains, documents, and metadata, the use of OSINT, the identification of potential threats, the notification module, and automated generation of PDF reports. The notification module includes asynchronous processing, exception handling, encrypted channels, event

logging, and REST API integrations, demonstrating a level of implementation beyond a purely conceptual model.

PyFormat is a practically significant result. The algorithm includes analysis, editing, configuration, and extraction, as well as deduplication. The experiment with a CSV file containing 2,974,336 rows and 49 fields demonstrates the processing of a large dataset, with the reported speed ranging from more than 1,000 to approximately 100,000 rows per second. The author states that more than 5 billion records were validated and converted to CSV format using PyFormat and that the algorithms were accelerated by more than 25 times during development. The comparison with MS Excel demonstrates an advantage over a conventional office product, but a stronger performance assessment would require comparison with specialized ETL/Big Data tools under identical conditions and with repeatable measurements.

The analysis of 3,723,514 real user passwords is one of the best quantitatively supported experiments. Password length, frequency, position, and character type are examined in order to identify dependencies related to the human factor in password creation. The geolocation analysis is implemented using the author's own data, which is an appropriate choice from an ethical perspective. EXIF/GPS data, IP-based and manually specified coordinates, Delaunay triangulation, and point weighting are used. A zone with an approximate radius of 500 meters is obtained, but the author himself notes that the result cannot be objectively verified. The weights assigned to the different sources of location data and their influence on the final result are also not sufficiently justified or quantitatively validated.

At the end of the chapter, a functional comparison between ASAID and existing systems is presented using the indicators "Yes/No/Limited". It demonstrates the broader intended functional scope of ASAID, but it does not constitute a quantitative benchmark in terms of latency, throughput, analytical accuracy, cost, and resource utilization. Overall, the material in Chapter Four demonstrates real software development and experimental work, although some architectural and analytical results remain prototype-based or demonstrative.

**5. Scientific and scientific-applied contributions of the dissertation**

The dissertation formulates three scientific-applied and two applied contributions. I accept the contributory nature of the work in substance, but I consider that the formulations could be refined so as to distinguish more clearly between novel results and review activities.

I regard the following as the most substantial scientific-applied results:

- The developed methodological sequence for transforming raw heterogeneous leaked data into verified, structured, and analytically usable datasets;
- The conceptual model of ASAID, which integrates collection, storage, processing, analytical modules, notification, and reporting;
- The adaptation and integration of different analytical procedures to the specific task of identifying potential threats.

As clearly expressed applied contributions, I accept the development and experimental use of PyFormat and the partial implementation of ASAID with functional modules for processing, analysis, notification, and report generation. The formulation "the main analyses applicable to leaked data have been reviewed and compared" is, in itself, primarily systematic in nature and I would not regard it as a strong independent scientific-applied contribution unless it is linked to the proposed methodology and architecture.

In summary, the contributions are predominantly scientific-applied and applied and are sufficient in nature for a dissertation in the doctoral programme "Computer Technologies in Engineering".

**6. Assessment of the doctoral candidate's personal contribution**

The presented material provides grounds for recognizing substantial personal involvement by the doctoral candidate. The text describes the author's own development of PyFormat, the construction and configuration of prototype ASAID components, the collection and processing of large information datasets, the development of analytical scripts, and the conduct of experiments. The location analysis uses the author's personal data, which further demonstrates direct involvement in the experimental work.

Of the three publications on the topic, one is single-authored and the other two are co-authored. On the basis of the dissertation text and the list of publications presented, I accept that the main results and contributions are to a substantial degree the personal work of the candidate.

**7. Assessment of the publications related to the dissertation**

Three publications related to the dissertation are presented and are thematically connected with the main directions of the research: a single-authored publication from 2022 on the use of dependencies derived from leaked data in password generation; a 2023 publication on geolocation authentication using leaked data; and a 2026 publication on an OSINT tool for geolocation analysis through coordinate triangulation, stated as accepted for publication. All are published in the Annual of the University of Mining and Geology "St. Ivan Rilski".

I positively assess the presence of a single-authored publication and the direct thematic relationship between the publications and the dissertation. The dissertation itself does not provide information on citations, indexing in international databases, or a separate bibliometric evaluation; therefore, such claims cannot be made solely on the basis of the submitted work. Compliance with the specific minimum national and institutional requirements should be established from the complete set of documents in the procedure.

**8. Use of the dissertation results in scientific and social practice. Practical applicability**

The results have a clear practical orientation. Prototype components have been implemented for searching compromised identifiers, processing and unifying data, analyzing passwords, locations, domains, documents, and metadata, automatic notification, and generation of PDF reports. The possibility of implementing ASAID in the cloud using Google Cloud was also examined, and a preliminary assessment was made of the required infrastructure and the associated monthly costs.

The practical value is demonstrated most strongly at the level of a research prototype and a proprietary software tool. No documents are presented for deployment of ASAID in a real organization, long-term operation, or direct economic impact. The result should therefore be assessed as an implemented prototype with potential for practical deployment rather than as a completed production system.

**9. Assessment of the conformity of the dissertation abstract with the requirements and the adequacy of its representation of the dissertation**

The dissertation abstract comprises 55 pages and has been prepared in accordance with academic requirements. Following comparison with the full dissertation, I consider that the abstract generally presents adequately the structure, stated tasks, architectural model, main experimental results,

contributions, and limitations of the research. The main results concerning PyFormat, password analysis, location analysis, functional modules, and comparison with existing systems are reflected consistently with the dissertation.

There are individual editorial inconsistencies, including a different number of tables stated – 10 in the abstract and 9 in the dissertation – as well as inaccuracies in the numbering and spelling of certain terms and names. These issues do not affect the substance of the research.

## 10. Remarks and recommendations

The following remarks and recommendations can be made regarding the dissertation:

- Different parts of the dissertation report different volumes of the data studied - "more than 3 billion" and "approximately 5 billion" records. There is no sufficiently clear distinction between the total number of records collected, verified, processed, and included in the individual experiments.
- The reported figure of approximately 95% verified data is not supported by a sufficiently formalized quantitative criterion. The approaches used for cross-checking and comparison with reference sources are described, but no methodology is presented that would allow this estimate to be reproduced unambiguously.
- The proposed cluster architecture of ASAID is implemented and investigated mainly in a virtualized environment with shared physical resources. Consequently, the degree of acceleration and scalability when moving to a real shared-nothing configuration has not been experimentally demonstrated, and some performance estimates are predictive in nature.
- The experimental results for PyFormat demonstrate operability with large information datasets, but the comparative performance assessment is limited. The use of MS Excel as the main comparison baseline does not allow a full evaluation against specialized ETL and Big Data processing tools.
- In the location-data analysis and the use of Delaunay triangulation, weights are assigned to different sources of coordinate information, but the method for determining these weights and their influence on the final result are not quantitatively validated. The author himself notes the limited possibility for objective verification of the results obtained.

The above remarks do not diminish the value of the dissertation and are primarily of a recommendatory nature, while also highlighting the substantial scope of the research work carried out.

## 11. Conclusion

The dissertation of Assist. Prof. MSc Eng. Ivan Lyubomirov Drankov addresses a relevant and practically significant problem in the field of information security, data analysis, and the design of scalable computer systems. The work follows a consistent logic - from analysis of existing solutions and formulation of an original classification, through a methodology for data preparation and verification and an architectural model, to prototype implementation and experimental studies. I regard the conceptual model of ASAID, the developed PyFormat software tool, the systematized procedure for processing large heterogeneous datasets, the functional analysis and notification modules, the analysis of a substantial dataset of real user passwords, and the automated generation of analytical reports as significant results.

Importantly, the work is not limited to a concept but includes software components that have actually been developed and investigated.

At the same time, the work has clearly defined limitations: cluster scaling has not been validated on independent physical nodes; some quantitative claims require stricter formalization; and certain experiments are demonstrative because of legal and ethical constraints. The author himself acknowledges a substantial part of these limitations and identifies directions for future development.

The presented results are predominantly scientific-applied and applied in nature and demonstrate the candidate's ability to independently formulate an engineering research problem, design an architectural solution, develop software, and conduct experimental studies with large information datasets.

I consider that the dissertation meets the requirements of the Law on the Development of Academic Staff in the Republic of Bulgaria and the relevant regulations. On the basis of the above, I give a positive assessment of the dissertation and confidently recommend that the esteemed Scientific Jury award Assist. Prof. MSc Eng. Ivan Lyubomirov Drankov the educational and scientific degree "Doctor".

Sofia

24 August 2026 Reviewer: .....

/Assoc. Prof. Georgi Zapryanov, PhD/