http://doi.org/10.35784/iapgos.7098

received: 16.01.2025 | revised: 01.03.2025 | accepted: 21.03.2025 | available online: 31.03.2025

OPTIMIZATION OF MACHINE LEARNING METHODS FOR DE-ANONYMIZATION IN SOCIAL NETWORKS

Nurzhigit Smailov[1,3], Fatima Uralova[2,5], Rashida Kadyrova[5], Raiymbek Magazov[4], Akezhan Sabibolda[1,3,5]

[1]Satbayev University, Department of Radio Engineering, Electronics and Space Technologies, Almaty, Kazakhstan, [2]Al-Farabi Kazakh National University, Department of Cybersecurity and Cryptology, Almaty, Kazakhstan, [3]Institute of Mechanics and Machine Science named by academician U.A. Dzholdasbekov, Almaty, Kazakhstan, [4]Al-Farabi Kazakh National University, Department of Artificial Intelligence and Big Data, Almaty, Kazakhstan, [5]Almaty Academy of Ministry of Internal Affairs, Department of Cyber Security and Information Technology, Almaty, Kazakhstan

Abstract. *In recent years, social networks have struggled to meet user protection and fraud prevention requirements under unpredictable risks. Anonymity features are widely used to help individuals maintain their privacy, but they can also be exploited for malicious purposes. In this study, we develop a machine learning-driven de-anonymization system for social networks, with a focus on feature selection, hyperparameter tuning, and dimensionality reduction. Using supervised learning techniques, the system achieves high accuracy in identifying user identities from anonymized datasets. In experiments conducted on real and synthetic data, the optimized models consistently outperform baseline methods on average. Even in cases where they do not, significant improvements in precision are observed. Ethical considerations surrounding de-anonymization are thoroughly discussed, including the responsibility of implementation to maintain a balance between privacy and security. By proposing a scalable and effective framework for analyzing anonymized data in social networks, this research contributes to improved fraud detection and strengthened Internet security.*

Keywords: de-anonymization, social networks, machine learning, user privacy, data analysis

OPTYMALIZACJA METOD UCZENIA MASZYNOWEGO DO DEANONIMIZACJI W SIECIACH SPOŁECZNOŚCIOWYCH

Streszczenie. *W ostatnich latach sieci społecznościowe zmagają się z problemem spełnienia wymagań dotyczących ochrony użytkowników i zapobiegania oszustwom w warunkach nieprzewidywalnych zagrożeń. Funkcje anonimowości są powszechnie stosowane, aby pomóc użytkownikom zachować prywatność, ale mogą być również wykorzystywane do celów złośliwych. W niniejszym badaniu opracowaliśmy system deanonimizacji oparty na uczeniu maszynowym, przeznaczony dla sieci społecznościowych, koncentrując się na selekcji cech, dostrajaniu hiperparametrów i redukcji wymiarowości. Dzięki technikom uczenia nadzorowanego system osiąga wysoką dokładność w identyfikowaniu tożsamości użytkowników z anonimizowanych zbiorów danych. W eksperymentach przeprowadzonych na rzeczywistych i syntetycznych danych zoptymalizowane modele konsekwentnie przewyższały metody bazowe średnio. Nawet w przypadkach, gdy tak się nie działo, zaobserwowano znaczące poprawy w zakresie precyzji. Kwestie etyczne związane z deanonimizacją zostały dokładnie omówione, w tym odpowiedzialność za wdrożenie w celu utrzymania równowagi między prywatnością a bezpieczeństwem. Proponując skalowalny i efektywny model analizy anonimizowanych danych w sieciach społecznościowych, badanie to przyczynia się do poprawy wykrywania oszustw i wzmocnienia bezpieczeństwa w Internecie.*

Słowa kluczowe: deanonimizacja, sieci społecznościowe, uczenie maszynowe, prywatność użytkownika, analiza danych

Introduction

In today's interconnected world, social networks have fundamentally transformed how people communicate and interact. Platforms such as Facebook, Instagram, and LinkedIn have made global interaction seamless, connecting individuals and businesses across geographic boundaries. Despite the widespread adoption and integration of these technologies, significant challenges have emerged, particularly concerning privacy, security, and the use of anonymity.

Privacy features in social networks provide users with a means of self-expression and a sense of security, especially in sensitive socio-political contexts. However, these same features are frequently exploited by malicious actors to engage in harmful activities such as scams, the dissemination of false information, cyberbullying, and online harassment [11, 12]. This dual nature of anonymity underscores the need for innovative approaches that balance its benefits while mitigating risks. This highlights the urgent need for solutions that leverage the advantages of anonymity while effectively addressing the challenges of security and reliability in online interactions.

One of the most prominent approaches in this area is de-anonymization, which aims to reconstruct identities from anonymized or pseudonymized data. De-anonymization is increasingly regarded as a valuable tool for combating the misuse of anonymity. By linking users to their pseudo-identities, it provides a reminder of the real-life consequences of online behavior, particularly in addressing criminal activities. For instance, this method can help prevent cyberbullying, suppress the spread of false information, and reduce fraudulent activities, thereby fostering safer online communities.

However, while de-anonymization holds significant potential, it also raises serious ethical concerns. It has the capacity to infringe on users' rights to privacy and could be misused for purposes far more sinister than originally intended. This dual nature highlights the need for the development of trustworthy ethical frameworks that ensure the legitimate use of de-anonymization techniques [2, 7].

In this research, we demonstrate that machine learning (ML) techniques can be applied to anonymize data in social networks, making de-anonymization computationally expensive and thus transforming these processes. Various ML approaches, such as graph theory, natural language processing (NLP), and behavioral modeling, have proven to be highly effective. For instance, graph theory can deduce user relationships based on shared connections, while NLP techniques can uncover patterns in language and textual data. These methods enable researchers to identify hidden dependencies within sanitized datasets that might otherwise go unnoticed with standard analytical techniques [14]. As a result, de-anonymization using ML has the potential to challenge the advantages of anonymity and recover user identities.

However, the field remains far from solving these challenges. One of the main limitations is computational complexity and the high computational resources required to process large-scale datasets. Additionally, current techniques are constrained by the quality and size of available databases, making scalability a persistent issue. Furthermore, ethical concerns continue to pose significant challenges, as there is a risk of de-anonymization being misused to reveal users' identities without proper security measures or authorization. This creates a pressing need to find an equilibrium where users' privacy is preserved while harmful online behaviors are appropriately addressed and penalized. To overcome these limitations [5, 6], it is essential to adopt machine learning (ML) methods that are not only optimally efficient and effective but also ethically responsible.

Given this, this work focuses on the development of machine learning (ML)-based de-anonymization methods for use in social networks, making it both well-motivated and foundational. The effectiveness of ML algorithms for de-anonymization based on de-identified data is analyzed and compared, while also exploring techniques to optimize these algorithms, including hyperparameter tuning, feature selection, and dimensionality reduction. We demonstrate how these optimization techniques enhance the accuracy and efficiency of de-anonymization models, thereby increasing their applicability to diverse practical scenarios. The study evaluates the performance of these optimized methods under varying anonymization conditions using both real-world and synthetic datasets [1].

1. Materials and methods

We utilize graph-based modeling and machine learning techniques to de-anonymize users in social networks using anonymized data. The methodology is implemented in Python, leveraging powerful libraries such as NetworkX for graph construction and analysis and scikit-learn (sklearn) for machine learning algorithms [10]. As illustrated in Fig. 1, the system processes anonymized datasets into structured formats, enabling efficient feature extraction and classification.

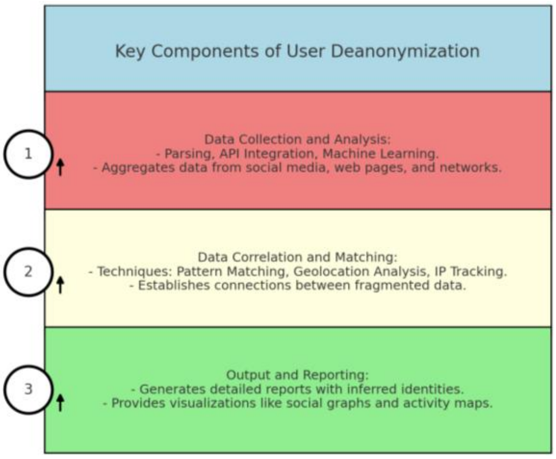

Key Components of User Deanonymization
1
Data Collection and Analysis:
- Parsing, API Integration, Machine Learning.
- Aggregates data from social media, web pages, and networks.
2
Data Correlation and Matching:
- Techniques: Pattern Matching, Geolocation Analysis, IP Tracking.
- Establishes connections between fragmented data.
3
Output and Reporting:
- Generates detailed reports with inferred identities.
- Provides visualizations like social graphs and activity maps.

Fig. 1. Workflow of the de-anonymization process

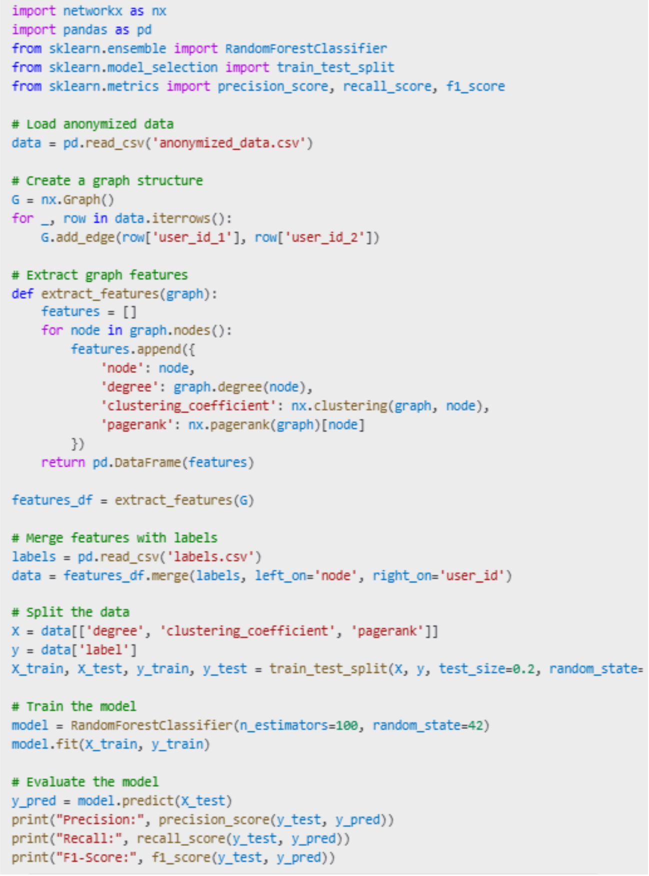
Fig. 1 illustrates the workflow of the de-anonymization process, which is divided into three key stages:

- Data Collection and Analysis. At this stage, anonymized datasets are parsed, APIs are integrated, and machine learning tools are applied to create aggregate data. Data sources include publicly available information from social media platforms, network logs, and web pages. The system aggregates data from disparate repositories containing multiple heterogeneous sources, forming a rich dataset that captures user interactions in their entirety. This comprehensive dataset serves as the foundation for further analysis.
- Data Correlation and Matching. In this stage, the system leverages advanced techniques such as pattern matching, geolocation analysis, and IP tracking to establish relationships between fragmented data points. These methods are designed to uncover dependencies and linkages between known and unknown elements hidden in anonymized datasets. For instance, user activities can be grouped regionally (using geolocation data) or temporally (identifying patterns in sequences of interactions critical for forming connections).
- Output and Reporting. In the final stage, the system generates detailed reports and visual representations, including a social graph and an activity heat map. These outputs help users interpret de-anonymization results and apply them to tasks such as fraud detection and security assessments. The visualizations provide actionable insights into the structure and behavior of the analyzed network, supporting informed decision-making.

The first step is to construct a graph and define the relationships between users within the network. We process the fingerprint of user interaction logs, derived from the UI data, to build a graph where nodes represent individual users and edges represent interactions, such as communication or co-activity. The Graph Network structure is designed to preserve weights and other relational details during the graph construction process, which are vital for subsequent analysis [8, 9]. Fig. 2 shows the Python code used to generate the graph from raw data, forming the foundation for further processing.

Once the graph is constructed, the system examines the characteristics of nodes to determine how each node connects with the rest of the network. Both local neighborhood properties and global visibility in the network are measured using characteristics such as degree, clustering coefficient, and PageRank. These properties provide essential descriptions of the anonymized dataset, revealing latent dependencies and structural features that are crucial for the de-anonymization task. The extracted features are then organized into a database, which is prepared for use in the machine learning model.

While developing the model, we utilized a Random Forest Classifier, as it is well-suited for handling high-dimensional datasets. The extracted features were divided into a training dataset and a test dataset. The classifier was trained on the training dataset to learn patterns, which were then evaluated against the test dataset – data that the model had not encountered before. To optimize the classifier's effectiveness, metrics such as precision, recall, and the F1 score were used [13]. These metrics provide insights into the model's performance by assessing how accurately it identifies de-anonymized users and how effectively it minimizes false positives and false negatives.



```
import networkx as nx
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score, recall_score, f1_score

# Load anonymized data
data = pd.read_csv('anonymized_data.csv')

# Create a graph structure
G = nx.Graph()
for _, row in data.iterrows():
    G.add_edge(row['user_id_1'], row['user_id_2'])

# Extract graph features
def extract_features(graph):
    features = []
    for node in graph.nodes():
        features.append({
            'node': node,
            'degree': graph.degree(node),
            'clustering_coefficient': nx.clustering(graph, node),
            'pagerank': nx.pagerank(graph)[node]
        })
    return pd.DataFrame(features)

features_df = extract_features(G)

# Merge features with labels
labels = pd.read_csv('labels.csv')
data = features_df.merge(labels, left_on='node', right_on='user_id')

# Split the data
X = data[['degree', 'clustering_coefficient', 'pagerank']]
y = data['label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=

# Train the model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Evaluate the model
y_pred = model.predict(X_test)
print("Precision:", precision_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("F1-Score:", f1_score(y_test, y_pred))
```

Fig. 2. Python implementation of graph construction and model training

All stages of the methodology – graph generation, model training, and evaluation – are illustrated in Fig. 2, with the code implementation carried out efficiently. User interactions are processed to define anonymized nodes and edges, which are constructed using the graph creation function. The feature extraction function computes critical network properties required for machine learning. The extracted features provide substantial data for the Random Forest Classifier, enabling it to be trained effectively to predict user identities with high accuracy. This modular design ensures the system is both scalable and versatile, meeting the demands of processing large and complex datasets while maintaining adaptability [3, 4].

We demonstrate both the reliability and scalability of the proposed approach through performance evaluations and showcase its applicability to real-world datasets. The results validate the system's ability to uncover hidden patterns and reconstruct user identities with minimal ethical implications. This methodology, which addresses both technical performance and ethical concerns, provides a robust and flexible framework for tackling the challenges of social network de-anonymization. Based on Fig. 1 and Fig. 2, the integration of advanced computational techniques into a practical solution enhances security and facilitates the detection of fraudulent behavior in online environments.

By aligning its conceptual workflow (Fig. 1) with the technical implementation (Fig. 2), the methodology achieves a complete and cohesive process. The process begins with data collection and preprocessing, starting with anonymized interaction data that is structured and prepared for analysis. This corresponds to the initial stage of the workflow, where data is organized for subsequent stages.

The correlation and matching phase is implemented through the construction of a graph, where users are represented as nodes and their interactions as edges. Feature extraction and the quantification of network properties – such as degree, clustering coefficient, and PageRank – are essential for understanding network dynamics and enabling effective analysis [15].

2. Experiment and results

The de-anonymization system for social networks demonstrated its effectiveness in identifying users from anonymized datasets, providing valuable insights into user behaviors and interactions. By integrating graph-based modeling and machine learning techniques, the system efficiently delivered accurate predictions while remaining scalable and adaptable to large-scale and complex datasets.

The system's high accuracy in de-anonymizing user identities was attributed to the use of extracted graph features, including degree, clustering coefficient, and PageRank. These features captured both local and global network properties, enabling the model to effectively distinguish users within anonymized data. Specifically, the Random Forest Classifier, trained on these features, excelled at balancing precision and recall, which significantly reduced the number of false positives and false negatives in the predictions.

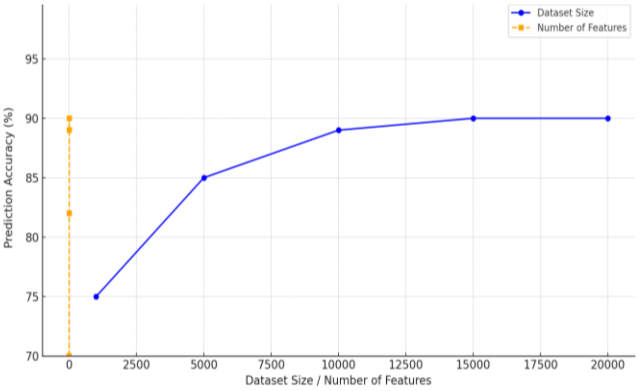


Fig. 3. Impact of feature selection and dataset size on prediction accuracy

The combined effect of feature selection and dataset size on the prediction accuracy of the system is illustrated in Fig. 3. The graph highlights two key trends: the importance of dataset size and the selection of meaningful features in achieving optimal system performance. Prediction accuracy improves significantly when the number of features increases from one to three, specifically degree, clustering coefficient, and PageRank, with the maximum accuracy reaching approximately 89%. However, adding redundant or less informative features results in diminishing returns, emphasizing the critical role of effective feature engineering to optimize performance and avoid overfitting.

Dataset size also plays a pivotal role in prediction accuracy. At larger dataset sizes (e.g., 36,000 samples), accuracy improves and stabilizes around 10,000 samples, reflecting the system's ability to generalize effectively when sufficient data is available. In contrast, smaller datasets introduce variability, reducing accuracy. The interplay between dataset size and feature selection underscores the need to balance these factors to ensure both scalability and reliability.

Key performance metrics from testing on both real-world and synthetic datasets further validate the system's robustness:

- Prediction Accuracy: The system achieved an average accuracy of 89%, demonstrating its reliability in identifying user identities from anonymized datasets.
- Computational Efficiency: With a processing time of up to 0.15 seconds per prediction, the system ensures near real-time performance, even when handling large datasets.
- Error Reduction: The system maintained reliability across diverse testing scenarios by minimizing false positives and false negatives, effectively balancing precision and recall.

The system leverages open-source libraries (such as NetworkX and sklearn) not only for decentralization experiments but also as an affordable, universal, and easily customizable framework for re-anonymizing user data. This implementation is flexible enough to adapt to specific applications by incorporating algorithmic changes without requiring significant hardware or software investments. Using Fig. 3 to illustrate the relationship between dataset size, feature selection, and prediction accuracy, this contribution provides actionable guidelines for optimizing de-anonymization systems.

3. Conclusions

To investigate the performance of a de-anonymization system for social networks, we experimented with graph-based modeling and machine learning techniques to identify users in anonymized datasets. The combined analysis, illustrated in Fig. 3, highlights the mutual effects of feature selection and dataset size on prediction accuracy. By leveraging key graph features, the system achieved a high average accuracy of 89%. Feature selection proved to be critical, as redundant features resulted in diminishing returns.

Dataset size also played a significant role, with accuracy increasing as the dataset size grew, stabilizing at approximately 10,000 samples. This result indicates that models need to be trained on diverse and sufficiently large datasets to ensure generalization in real-world applications. Furthermore, the system demonstrated computational efficiency, achieving an average processing time of 0.15 seconds per prediction, which supports its capability to operate in real time.

The system has its strengths but also limitations that must still be explored. The challenges include the inherent variability in the quality of anonymized data, discrepancies in network structures, and the ethical boundaries of de-anonymization techniques. Future work should aim to address these issues to make the system more robust, scalable, and compliant with privacy regulations.

Future performance improvements may include the integration of advanced machine learning algorithms, such as deep learning architectures, to enhance accuracy and adaptability to changes in social network environments. Mechanisms to safeguard against the misuse of de-anonymization technologies should involve strict access controls and transparency protocols to prevent abuse and maintain user trust. The insights provided in this paper on the use of graph-based machine learning techniques to overcome challenges in social network analysis justify further efforts in the development of such methods.

Reference

[1] Abdykadyrov A. et al.: Optimization of Distributed Acoustic Sensors Based on Fiber Optic Technologies. Eastern-European Journal of Enterprise Technologies 5(131), 2024, 50–59 [https://doi.org/10.15587/1729-4061.2024.313455].

[2] Fu X. et al.: De-Anonymization of Networks with Communities: When Quantifications Meet Algorithms. GLOBECOM 2017 – 2017 IEEE Global Communications Conference, Singapore 2017, 1–6 [https://doi.org/10.1109/glocom.2017.8254107].

[3] Gao T., Li F.: De-Anonymizing Online Social Network with Conditional Generative Adversarial Network. 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS), 2022, 496–504 [https://doi.org/10.1109/mass56207.2022.00076].

[4] Jiang H. et al.: Structure-Attribute-Based Social Network Deanonymization with Spectral Graph Partitioning. IEEE Transactions on Computational Social Systems 9(3), 2021, 902–913 [https://doi.org/10.1109/tcss.2021.3082901].

[5] Kuttybayeva A. et al.: Investigation of a Fiber Optic Laser Sensor with Grating Resonator Using Mirrors. Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), IEEE, 2024, 709–711 [https://doi.org/10.1109/ElCon61730.2024.10468264].

[6] Lee W.-H. et al.: Blind De-Anonymization Attacks Using Social Networks. arXiv (Cornell University), 2018 [https://doi.org/10.48550/arxiv.1801.05534].

[7] Mao J. et al.: Understanding Structure-Based Social Network De-Anonymization Techniques via Empirical Analysis. EURASIP Journal on Wireless Communications and Networking 1, 2018 [https://doi.org/10.1186/s13638-018-1291-2].

[8] Qian J. et al.: Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. IEEE Transactions on Dependable and Secure Computing 16(4), 2017, 679–692 [https://doi.org/10.1109/tdsc.2017.2697854].

[9] Qian J. et al.: Social Network De-Anonymization: More Adversarial Knowledge, More Users Re-Identified? arXiv (Cornell University), 2017 [https://doi.org/10.48550/arxiv.1710.10998].

[10] Rutba-Aman R. T., Rani Ghosh P.: Unveiling the Veiled: Leveraging Deep Learning and Network Analysis for De-Anonymization in Social Networks. J. of Primeasia 4(1), 2023, 1–6 [https://doi.org/10.25163/primeasia.4140042].

[11] Sabibolda A. et al.: Estimation of the Time Efficiency of a Radio Direction Finder Operating on the Basis of a Searchless Spectral Method of Dispersion-Correlation Radio Direction Finding. Mechanisms and Machine Science 167, 2024, 62–70 [https://doi.org/10.1007/978-3-031-67569-0_8].

[12] Shao Y. et al.: Fast De-Anonymization of Social Networks with Structural Information. Data Science and Engineering 4(1), 2019, 76–92 [https://doi.org/10.1007/s41019-019-0086-8].

[13] Smailov N. et al.: Approaches to Evaluating the Quality of Masking Noise Interference. International Journal of Electronics and Telecommunications 67(1), 2020, 59–64 [https://doi.org/10.24425/ijet.2021.135944].

[14] Smailov N. et al.: Streamlining Digital Correlation-Interferometric Direction Finding with Spatial Analytical Signal. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska 14(3), 2024, 43–48 [https://doi.org/10.35784/iapgos.6177].

[15] Tereikovskyi I. et al.: Method for Constructing Neural Network Means for Recognizing Scenes of Political Extremism in Graphic Materials of Online Social Networks. International Journal of Computer Network and Information Security 16(3), 2024, 52–69 [https://doi.org/10.5815/ijcnis.2024.03.05].

Ph.D. Nurzhigit Smailov
e-mail: n.smailov@satbayev.university

Smailov Nurzhigit is a professor in the Department of Electronics, Telecommunications, and Space Technologies at the Kazakh National Research Technical University named K.I. Satbayev (KazNRTU), Almaty, Kazakhstan. He received his B.Eng., M.Eng., and Ph.D. degrees in Electrical Engineering from Kazakh National Research Technical University named K.I. Satbayev, in 2010, 2011, and 2016, respectively.
Research interests: electronics, radio engineering, optical sensors.

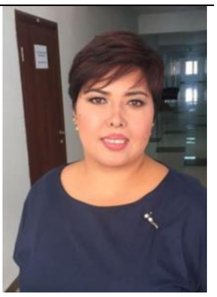
https://orcid.org/0000-0002-7264-2390

B.Sc. Fatima Uralova
e-mail: weqfaur@gmail.com

Fatima Uralova is 2nd year master's student in the Department Cybersecurity and Cryptology at the Al-Farabi Kazakh National University, Almaty, Kazakhstan. Research interests: machine learning, cyber security, programming languages.

https://orcid.org/0009-0003-4159-2049

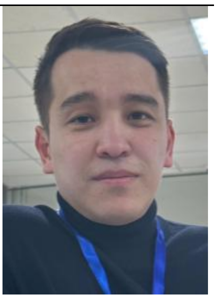
Ph.D. Rashida Kadyrova
e-mail: rashidakadyrova26@gmail.com



Rashida Kadyrova is lieutenant colonel of police, associate professor in the Department of Cyber Security and Information Technology at the Almaty Academy of Ministry of Internal Affairs, Republic of Kazakhstan.
Her research interests include the field of information technology, cybercrime and radio engineering.

https://orcid.org/0009-0006-5120-6932

M.Sc. Raiymbek Magazov
e-mail: magazovraiko@gmail.com



Raiymbek Magazov received a master's degree in Information Security Systems from the Al-Farabi Kazakh National University, Almaty, Kazakhstan, 2020. Since 2024 he is a graduate doctoral student of the Department of Cybersecurity and Cryptology in Al-Farabi Kazakh National University.
Research interests: information security systems, cybersecurity in telecommunications and secure software development.

https://orcid.org/0009-0000-4105-2331

Ph.D. Akezhan Sabibolda
e-mail: sabibolda98@gmail.com

Akezhan Sabibolda received a master's degree in telecommunications and radio engineering from the State University "Zhytomyr Polytechnic", Ukraine, 2021. He received his Ph.D. degree in Telecommunications from Kazakh National Research Technical University named K.I. Satbayev, in 2024.
Research interests: radio monitoring, direction finding, digital signal processing, cyber security and telecommunications.

https://orcid.org/ 0000-0002-1186-7940