ALESIA CHERNIKOVA

Postdoctoral Research Associate, Northeastern University, Boston, USA
a.chernikova@northeastern.edu — 1-781-350-0139 — LinkedIn — achernikova.github.io — Google Scholar — GitHub

SUMMARY

AI researcher and engineer specializing in geometric deep learning and AI reliability – adversarial robustness, uncertainty quantification, generalization, mechanistic interpretability, alignment. Built scalable graph analytics pipelines for AWS cloud activity tracing and anomaly detection. Ten years of research experience and four years of industry software engineering. Strong in Python/Scala, PyTorch/JAX/TensorFlow, and Spark.

SELECTED HIGHLIGHTS

- **Uncertainty Quantification and Theoretical Guarantees in Message-Passing Neural Networks (MPNNs)**: Built the first end-to-end theory of uncertainty-aware MPNNs, connecting moment propagation, adversarial robustness, and generalization. Derived analytic uncertainty propagation rules for MPNNs, including principled approximations for nonlinear activations. Introduced node-level certified robustness guarantees linking predictive uncertainty to adversarial safety. Defined Feature Convolution Distance (FCDp), enabling generalization bounds for graph learning with noisy features. [AISTATS 2026, NeurIPS 2025 workshop: NPGML.]
- **Mechanistic Interpretability of Large Language Models (LLMs)**: Developing a fully automated framework for circuit extraction in LLMs using Cross-Layer Transcoders and attribution graphs, combining importance-weighted pruning with a flow-based method to identify compact computational circuits for diagnosing alignment failures and enabling targeted interventions to mitigate misalignment. [Current work in progress (GitHub).]
- **AI Security in Safety-Critical Domains**: Developed FENCE, a framework for feasible adversarial evasion under real-world constraints that enables gradient-based attacks on structured and tabular data. Demonstrated practical vulnerabilities in deployed systems for network intrusion detection and malicious domain classification, and showed that adversarial training can substantially improve resilience. Analyzed the robustness of deep learning in autonomous driving, introducing the first adversarial attacks on regression-based steering angle prediction, where imperceptible perturbations induce targeted control failures and large performance degradation (up to 69×). [ACM TOPS 2022 (GitHub), IEEE S&P 2019 workshop: SafeThings (GitHub).]
- **Scalable Security Graph Analytics of AWS cloud**: Designed and implemented a scalable graph-based system for tracing AWS cloud activity, modeling infrastructure events as a heterogeneous graph and enabling lateral movement detection using Bayesian inference and network science–based methods. [AWS internal conference 2021.]

EDUCATION

Northeastern University, Boston, USA, Doctor of Philosophy in Computer Science 09/2017 — 04/2024
Advisor: Alina Oprea, "Cyber networks resilience against adversarial attacks". GPA: 3.9/4.00

Belarusian State University, Minsk, Belarus, Bachelor of Science in Applied Mathematics 09/2009 — 06/2014
Advisor: Vladimir Malugin, "Development of risk management algorithms based on derivatives contracts". GPA: 3.8/4.00

ACADEMIC EXPERIENCE

Northeastern University, RADLAB, DK-Lab Boston, USA
Postdoctoral Research Associate 05/2024 — Present

- Developed a framework to quantify uncertainty in Message Passing Neural Networks (MPNNs), derived probabilistic certified adversarial robustness radii of MPNNs for the node classification task, and introduced a novel covering number-based generalization bound for the node classification task in MPNNs under uncertainty in node features.
- Develop a fully automated circuit extraction pipeline for LLMs using Cross-Layer Transcoders and attribution graphs to isolate compact circuits that enable systematic diagnosis of alignment failures and support targeted interventions.
- Design hyperbolic geometry–based sparse Deep Neural Networks (DNNs) architectures and alternative training pipelines, achieving efficiency, interpretability, and generalization of DNNs.

Northeastern University, NDS2 Lab Boston, USA
Research Assistant 09/2017 — 04/2024

- Introduced a new compartmental model of epidemiology to represent self-propagating malware (SPM) propagation in networks using real-world WannaCry traces. Rigorously studied the characteristics of the SPM propagation process under the homogeneous mixing assumption and on arbitrary networks.
- Designed new defense algorithms leveraging graph theory and extensively tested the behavior of existing defense techniques to improve the network robustness of large enterprise networks in the face of SPM.

- Developed a novel optimization-based framework for evasion attack algorithms that preserve possible feature dependencies to evaluate the robustness of deep learning models in constrained environments such as cybersecurity or healthcare. Evaluated the success of existing defense algorithms against the proposed attack methodology.
- Demonstrated first evasion attacks against classification and regression deep learning models in the domain of self-driving cars.

Amazon Web Services (AWS) Boston, USA
Research Scientist Intern 05/2020 — 08/2020, 05/2021 — 09/2021

- Created a scalable algorithm for tracing the activity in the AWS cloud represented as a heterogeneous graph to allow further research based on AWS cloud activity data.
- Developed the methodology for lateral movement detection in the AWS cloud environment using Bayesian statistics and network science perspectives.

Belarusian State University (BSU) Minsk, Belarus
Undergraduate Research Assistant 01/2012 — 12/2013

- Participated in a research project to create the methodology for estimating credit rankings of enterprises using clustering and factor analysis.
- Led the credit rankings estimation project for the building enterprise sector.
- Collaborated in developing the package for automated calculation of credit scores based on the proposed credit rankings evaluation methodology.
- Developed novel methodologies for hedging strategies using futures and interest-rate swap contracts.

PROFESSIONAL EXPERIENCE

IBA Group Minsk, Belarus
Senior Software Engineer 11/2013 — 07/2017

- Participated in the development of a large-scale IBM GSAR web portal and maintained its performance.
- Assisted the software architect with the efficiency and usability improvement of the portal.

TEACHING EXPERIENCE

Northeastern University Boston, USA
Invited Lectures for CS 7332: Machine Learning with Graphs 09/2024 — 12/2024, 09/2025 — 12/2025

- Designed and held lectures for a graduate-level course in Graph Machine Learning.

Northeastern University Boston, USA
Teaching Assistant for CS4100: Artificial Intelligence 09/2022 — 12/2022, 09/2023 — 12/2023, 01/2024 — 04/2024

- Designed and held lectures for undergraduate and graduate students.
- Held weekly office hours to answer questions, provide support, and review course material with students.
- Graded assignments, exams, and research projects.
- Assisted professor with homework and exam preparation, proctored the exams.
- Advised students regarding research projects.

PUBLICATIONS

Robustness and Generalization in Uncertainty-Aware Message Passing Neural Networks. AISTATS 2026
A. Chernikova, M. Laber, N. Sabhahit, T. Eliassi-Rad
Uncertainty-Aware Message Passing Neural Networks. New Perspectives in Advancing Graph Machine Learning,
A. Chernikova, M. Laber, N. Sabhahit, T. Eliassi-Rad NeurIPS 2025
Modeling Self-Propagating Malware with Epidemiological Models. Applied Network Science 2023
A. Chernikova, N. Gozzi, S. Boboila, N. Perra, T. Eliassi-Rad, and A. Oprea.
Cyber Network Resilience against Self-Propagating Malware Attacks. ESORICS 2022
A. Chernikova, N. Gozzi, S. Boboila, N. Perra, P. Angadi, J. Loughner, M. Wilden, T. Eliassi-Rad, and A. Oprea.
Fence: Feasible Evasion Attacks on Neural Networks in Constrained Environments. ACM TOPS 2022
A. Chernikova and A. Oprea.
Are Self-Driving Cars Secure? SafeThings IEEE S&P 2019
Evasion Attacks against Deep Neural Networks for Steering Angle Prediction.
A. Chernikova, A. Oprea, C. Nita-Rotaru and BG. Kim
Hedging Algorithms Based on Interest-rate Swaps. BSU Conference 2013
A. Chernikova and V. Malugin.

AWARDS

IAIFI Summer School Best Project Award 2025
IEEE S&P and GREPSEC Travel Grant 2019
Khoury College of Computer Science Fellowship 2017 — 2018
National Bank of the Republic of Belarus Merit Scholarship 2013 — 2014
Belarusian State University Excellence Merit Scholarship 2012 — 2014

TALKS

Uncertainty-Aware Message Passing Neural Networks. NPGML, NeurIPS, 2025
Circuit Tracing in Large Language Models. Network Science Institute, 2025
Modeling Self-propagating Malware with Compartmental Models of Epidemiology. JMM, 2025
Cybernetwork Resilience against Self-Propagating Malware Attacks. Network Science Institute, 2024
Cybernetwork Resilience against Self-Propagating Malware Attacks. DoD SERDP Workshop, 2024
Towards Resilient Cybernetworks against Adversarial Attacks. Amazon Web Services, 2023
Cybernetwork Resilience against Self-Propagating Malware Attacks. ESORICS, 2022
Feasible Evasion Attacks in Constrained Environments. CRA Seminar, 2022
Graph-based Statistical Detection of Anomalous Role Assumption Events. Amazon Web Services, 2020
Feasible Evasion Attacks on Neural Networks in Constrained Environments. ARL Meeting, 2020
Evasion Attacks against Deep Neural Networks for Steering Angle Prediction. SafeThings IEEE S&P, 2019

SERVICE

Reviewer — ACM TOPS, IEEE Transactions on Privacy
Technical Program Committee — AISTATS'26, IEEE S&P'26, IEEE S&P'25, IEEE MILCOM AI for Cyber'23

SKILLS

- **Programming:** Python, Java, Scala, Javascript, C/C++
- **Frameworks and Libraries:** JAX, PyTorch, Tensorflow, Keras, Spark