CS8803-O23 Modern Internet Research Methods

Course Syllabus for Summer 2026

Delivery: 100% Web-Based, Asynchronous

Instructor Information

Course Instructor: Dr. Maria Konte, mkonte@gatech.edu
Head TAs: Cody Tessler, ctessler3@gatech.edu; Anita Rao (arao338@gatech.edu)

About This Course

Welcome! This is a research-oriented course that covers new developments in Internet measurement techniques, with an emphasis on topics related to reliability, freedom, and security of modern Internet platforms.

The goals of this course are to:

a. Explore new research topics in the modern Internet interdisciplinary research areas.
b. Familiarize and experiment with techniques, tools, platforms and datasets.
c. Develop new research ideas and deliver an academic research paper. Use the course material as a starting point to brainstorm new research ideas and select a topic of interest. Perform the entire cycle from selecting a research topic, focusing on a specific research question, following through (e.g., data collection and analysis, system design and evaluation, etc.), and finally delivering the results through an academic paper.

Areas covered through the course

The topics the course discusses span three areas:

1. **Measurement Techniques for Internet and Cybersecurity Analytics**: Techniques to map and study the Internet host population along with the services it offers, and how these techniques can be leveraged for different applications, e.g., Internet infrastructure resilience, or cybersecurity analytics.

   Topics covered:

   - **Identifying and studying the "live" Internet host population at scale**. Passive and active scanning techniques for reliably identifying which IPv4 addresses correspond to active, reachable devices on the Internet.
   - **IP space utilization inference**: Given IPv4 exhaustion, how can we more precisely infer which portions of the address space are actively utilized

versus assigned-but-idle or completely unassigned? Accurate measurements have important implications for resource management and policy.

- **Service prediction across ports**: Techniques to predict what services (e.g., HTTP, SSH) are running on any port—not just standard ones—using machine learning models, improve our visibility about how services are deployed, even when they don't follow expected port selections.
- **Spoofed traffic detection.** Measurement techniques to detect spoofed traffic that might interfere with scanning campaigns, data collection and measurements.
- **Internet infrastructure hijacking.** Techniques that can be (ab)used to overtake control of the Internet infrastructure (DNS & BGP hijacking). Techniques to detect infrastructure hijacking.
- **The certificate ecosystem.** A brief look into Web management. A measurement approach to identify **stale or revoked certificates** and how they may be abused.
- **Internet voting** systems. An example technique to identify possible risks.

2. **Measurement Techniques for Studying Blocked Internet Access**: Techniques currently used to block access and measurement techniques used to detect when blocking is taking place.

   Topics covered:

   - What is censorship? Reviewing diverse **techniques used to enforce censorship**.
   - How are **censorship observatories designed**? An example observatory. Leveraging public observatories for longitudinal analysis.
   - Measurement techniques to **locate censorship devices.**
   - From rules to **machine learning based approaches** for detecting censorship **at scale**.

3. **Measurement Techniques for Understanding Abuse on Online Platforms**:

   Techniques used to map out the current landscape of abuse (toxicity, harassment, misinformation, etc.). Identify abusive behaviors and study abusive accounts on online platforms. Techniques that leverage social platforms for early warning.

   Topics covered:

   3A. Abuse on social platforms: hate, toxicity and harassment.

   - Classifying content related to hate, harassment and toxicity.
   - Detecting **toxic accounts and studying their trends and patterns.**

3B. How entities abuse online platforms.

- Studying **sock puppet accounts.**
- **Online cybercrime** communities and approaches to understanding them.
- **Mining the GitHub platform** to identify suspicious repositories.

3C. False information spread on online platforms and the supporting Internet infrastructure.

- Identifying **false information** on online platforms.
- **Measuring relationships** among websites that spread misinformation.
- Studying the **ecosystem of Internet infrastructure that supports** misinformation websites.

4. **Sustainable research and ethics**: Finally, the course covers topics related to ethics guidelines when performing large-scale Internet measurements and discusses elements of sustainable research, such as transparency and reproducibility.

   Topics covered:

   - **Reproducibility and replicability** for experimental networking research
   - **Ethical frameworks** for Internet measurement studies

Research Project Format (or "Avenues to Approach a Research Topic")

The research areas we cover include lectures, papers, and presentations. This list only serves as a starting point. The students are welcome and highly encouraged to branch out and explore from there, cutting across traditional boundaries.

Also, there are different "avenues" to explore a research question. The students are encouraged to shape their projects based on their **background, interests, and goals**.

Common approaches include:

- **Literature Review**: Conduct a systematic review of existing research – typically does not involve coding.
- **Survey Study**: Design and distribute a questionnaire and analyze the results to uncover trends or patterns.
- **Data Analysis**: Collect a dataset or use a public one to perform exploratory or in-depth analysis.
- **Learning Techniques**: Apply, evaluate, or even design an AI/ML method for a dataset related to a topic.

- **System Design**: Build and evaluate a system (e.g., a tool, pipeline, or framework) that tackles a specific problem.
- **Replication Study**: Reproduce and reassess results from a previously published paper—this could include publishing new datasets, re-running experiments, or testing under different conditions.
- **Prototype & White Paper**: Design and build a tool through a prototype that shows the core functionality, and write a white paper (typically short) that explains the main technical aspects.

Lectures Format

The instructor presents weekly prerecorded lectures, each of which covers a topic along with associated techniques, platforms, or datasets (see course calendar below). The lectures are delivered across 15 modules. Through each module, the instructor presents 2-4 research papers delivered over 2-4 short videos. In parallel, the students are assigned milestones that guide them towards the completion of their research project and paper (see the assignments section below).

Course Learning Objectives

The learning objectives of this course are to:

1. Describe the current state of research in the intersection of Internet measurements and cybersecurity. Specifically, you will learn about modern topics of broad and current interest related to the risks that the Internet infrastructure faces (e.g. Internet infrastructure hijacking), Internet censorship, abuse and entities on social platforms, web trust management, the ecosystem of false information on the web. Finally, you will learn about how to perform Internet measurements using ethical guidelines, and principles of sustainable research (e.g. replicability).
2. Describe a plethora of passive and active measurement techniques, data collection and analysis approaches for each of the above topics.
3. Demonstrate the ability to apply the learned techniques to different or new research questions.
4. Demonstrate the ability to put together a research project; from identifying a broad idea, to specifying a well-defined research question, outlining and executing a research approach to address it.
5. Demonstrate the ability to transfer a research project into an academic paper, and deliver a presentation of the paper.

Course Materials

There are no required books for this course. The lectures are based on academic papers.

Office Hours

See Ed Discussions for details.

Course Assignments and Grading

As a research-oriented course, the main component and focus of the course is a semester-long research project. Each student, either individually or in a group, will work on a research project that will run throughout the semester. More specifically, the assignments and their weights in the final grade are provided below. See Canvas for a rubric.

- **Acknowledge the course policies [0%]:** The course policies must be acknowledged. Please reach out to the instructor if you have questions about the class policies.
- **Research Interests questionnaire [1%]:** Students will complete a short questionnaire on Canvas and cross-post to Ed. This helps students find peers with aligned interests, complementary skills, and compatible research goals for forming groups.
- **Weekly Check-in Log & Meetings [14%]:** These check-ins serve to support progress, identify challenges, and ensure accountability.
    - Beginning in Week 2, each group meets with the instructional team every week.
    - You must meet with the professor the week after submitting your proposal and after each milestone. In other weeks, you may meet with any member of the course staff for feedback.
    - Each student submits the same brief weekly log in on:
        - Canvas for grading
        - Ed to share with their classmates how the progress is progressing
    - This log summarizes their individual contributions, challenges, and plans. These reflections guide the short weekly check-in meeting.
    - Weekly meeting attendance, participation, and thoughtful responses are required to receive full credit.
- **Brainstorming assignments (I and II) [6%]:** Each student begins by working on a brainstorming assignment, which later evolves into the project proposal to guide the final project. Towards this goal, and as intermediate steps, each student will submit

two brainstorming assignments that summarize the group's progress towards the proposal. The brainstorming assignments reflect progress as the group refines the project idea and gets ready for the proposal.

- **Research Problem and Related Work [5%]:** Each group starts a running draft of the paper and describes the main research problem they will work on, the paper's main contributions (typically a list of three main contributions), and the related work section.
- **GT GitHub repository and Overleaf Project [1%]:** Each group starts a GT GitHub repository and a GT Overleaf Project and shares the link with the instructional team.
- **Project Proposal [5%]**: Each group expands its running draft to include all sections of the paper in a skeleton format. The project proposal will explain the problem and approach in detail. Also, the proposal will include a schedule that outlines how the group will distribute the work throughout the semester and among the team members. Therefore, each team member has the tasks they are working on. Each group will split their work into milestones (for example, milestones can include: data collection, data analysis, system design, system implementation, system evaluation).

  After the project proposal deadline, each group is set in terms of team members, specific research problem, and approach. No further changes are allowed after this point.
- **Identify the target conference [1%]:** Submit the URL of the conference where you plan to submit, along with information on deadlines and formatting, or the type of submission.
- **Research Milestones [36% equally distributed]:** Teams will submit a PDF of the project's current state based on the milestones stated in the project proposal. The due dates for each milestone are listed in the schedule below.
- **Lecture Quizzes [5%]:** Students will complete a quiz on each lecture. The quizzes are "open book" and open resources. *Each module has an associated quiz, and all quizzes are due at the end of Week 15.*
- **Final Project code [4%]:** Each group submits the code for the project. Each group is required to use Georgia Tech's GitHub repository to host the project code throughout the course.
- **Final Paper presentation [6%]:** Each group will prepare and record a 15–20-minute presentation on their project.
- **Final Paper [6%]:** Each group will write a 10- to 12-page final report written in academic style format. Each group is required to use Georgia Tech's Overleaf to host the paper throughout the course.

- **Class Roundtable Discussions [10%]:** In the first two weeks of the course, our roundtable discussions will be to share our research interests on the course's Ed Discussion forum. Please see the Ed Discussion thread and write a comment there to describe your research interests to your classmates. From Week 3 onward, at least one student will be assigned to give a 5- to 10-minute presentation on the problem they are working on (e.g., the specific problem, a paper they recently read). The rest of the students, who are not assigned to present, comment on the presentation in a constructive manner, with follow-up questions. Each student is assigned to give a short presentation once throughout the semester. The instructor will coordinate with each student about which week to do their short presentation.
- **Extra Credit I: Results dissemination [2% extra credit]:** Each group is encouraged to disseminate their project either with a general high-level description or a more detailed description of the results. Examples can include: **1)** Setting-up a webpage on sites.gatech.edu, (setting up the website takes only a few minutes!) and include your title and abstract, **2)** Submit the link to your presentation to the OMSCS conference and showcase, **3)** Create a public GitHub repository hosting your code and instructions how to run your scripts, **4)** Reach out to the instructors if you have other ideas, e.g. an interactive environment where others can run your scripts.
- **Extra Credit II: Course surveys [1%]:** Throughout the semester, the instructional team will send out three surveys to provide course feedback.
- **Extra Credit III: Course Contributions [up to 10% extra credit, at the discretion of the course staff]:** As their research progresses, groups may develop artifacts that are beneficial to other researchers or future iterations of the course. Examples include (but are not limited to) data collection scripts, documented processes for PACE, benchmarks, or well-motivated alternative research directions.

  Groups may submit such contributions to the course staff, along with a 1- to 2-page written summary describing the artifact, its purpose, and its potential value to others. Extra credit will be awarded based on the quality, originality, and usefulness of the contribution.

  Exceptional and substantive contributions may also be acknowledged in a Course Contributions section of the syllabus.

Forming a Team, Team Size, and Proposed Work

Why work within a group? As is highly likely, you will confirm for yourself that writing an academic paper is different than putting together an end-of-semester class report, and it greatly benefits from teamwork. **The goal of this course is to guide you on how to write an academic paper on a topic that you are passionate about and to get it published.**

If we take a look at Google Scholar to identify our favorite papers, we will notice that, in practice, **almost no academic paper is written by a single author!** And there are several good reasons for that, including: 1) Performing the entire research cycle from conceiving an idea to narrowing down to a specific topic, and delivering the results through an academic paper is a very rewarding and time-consuming process. 2) Brainstorming, interacting with others, problem-solving, and pushing through can make your paper so much better and increase the chances that your paper will be accepted in top academic conferences.

How do we go about forming groups? Students will fill out a research questionnaire and cross-post it on Ed. Using this, students are asked to form groups of two to four peers with aligned interests, complementary skills, and compatible research goals.

What about team size? Please form groups of two to four students.

How can groups communicate? Groups can utilize any platform they wish to coordinate amongst themselves. All students will be added to a dedicated Slack channel on Georgia Tech's Slack workspace, and in addition, they will be pointed to a dedicated Ed Thread for their project to share their updates.

Grading Scale

The final grade will be assigned as a letter grade according to the following scale:

- A 90-100%
- B 80-89%
- C 70-79%
- D 60-69%
- F 0-59%

Course Prerequisites

The course is geared towards students who are interested in pursuing a research project and writing an academic paper.

This is a research-oriented course that intersects topics in Internet protocols, computer networks, cybersecurity and data analysis. Having taken courses in topics related to systems, ML/AI, data visualizations, data structures, algorithms, computer architecture is a plus, since the student will be able to leverage their background in these areas to pursue a research project and write an academic paper. The course will not cover undergraduate material typically covered in undergraduate networking, cybersecurity or data analysis courses. The students are expected to code in Python (or a language of their choice) at an intermediate level (e.g., comfortably using object-oriented programming, data structures, control structures, etc., as well as testing and debugging tools/strategies).

In lieu of a readiness questionnaire, prospective students are expected to be comfortable with and/or passionate about:

- Reading and understanding the paper "*An Open Platform to Teach How the Internet Practically Works".*
- Defining their research questions/ideas and, therefore, working with open-ended projects rather than predefined assignments.
- Student-led projects that require more autonomy and taking ownership of the work and the progress/pace.
- Working with projects that require coding skills, as well as technical writing and presentation skills.
- Receiving peer-to-peer feedback.
- Working in a group of students with multidisciplinary backgrounds.

Course Calendar

Unless otherwise specified, all deliverables and check-in logs are due on Sundays at 11:59 PM AOE.

The initial post for the **class roundtable discussions** is due on Thursday at 11:59 PM AOE, and the responses are due on Sunday at 11:59 PM AOE. The schedule for presentations will be posted on Ed Discussion.

The **Modules** can be completed in any order. A recommended pacing is provided. Each module has an associated quiz, and all quizzes are due at the end of the semester.

Week	Dates	Module(s)	Key Deliverables
1	May 18 - May 24	1	Research Interests Questionnaire Brainstorming I Course Policy Acknowledgement
2	May 25 - May 31	2-3	Brainstorming II Group Formation
3	Jun 1 - Jun 7	4	Research Problem and Related Work GitHub and Overleaf Setup
4	Jun 8 - Jun 14	5-6	Proposal
5	Jun 15 - Jun 21	7-8	Identify Target Conferences Meet with Professor
6	Jun 22 - Jun 28	9-10	Course Survey I
7	Jun 29 - Jul 5	11-12	Milestone I
8	Jul 6 - Jul 12	13-14	Course Survey II Meet with Professor
9	Jul 13 - Jul 19	15	Milestone II
10	Jul 20 - Jul 26		Course Survey III Meet with Professor
11	Jul 27 - Aug 2		Submit final paper, code, and presentation CIOS Survey Extra Credit: Results dissemination
	July 27 (Mon)	-	Final Day of Instruction
	August 6 (Thu)	-	Quizzes Due End of Term
	August 10 (Mon)		Grade Submission Deadline
	August 11 (Tue)		Final grades available after 6:00 PM ET

Module Summaries

Legend (modules → titles):

1. Course Overview and Course Overview and Reading, Writing & Presenting Papers
2. Surveying the Internet Address Space (I)
3. Surveying the Internet Address Space (II)
4. Overtaking the Internet Infrastructure Control
5. Internet Censorship [I]
6. Internet Censorship [II]
7. Web and Trust Management
8. Measurements & Voting Systems
9. Abuse on Social Platforms
10. Entities on Social Platforms
11. False Information on Web & Social Media
12. The False Information Ecosystem
13. Online Platforms as a Vantage Point to Study Cybercrime Communities
14. Ethics in Internet Measurements
15. Sustainable Research: Transparency & Reproducibility

Module 1: Course Intro & Crash Review: "How the Internet Works" and Reading, Writing, and Presenting Papers

In this module, we first provide a walkthrough of the course map and the topics we will discuss through the modules, as well as the learning objectives of the course. As a quick review of Internet protocols and how the Internet works, we discuss an open platform that offers opportunities for experimentation.

Next, we look at a technique for reading academic papers, the basic "ingredients" of a well-written paper, and some writing and presentation tips.

Topics:

- What is the course about? Learning goals
- Overview of course topics and techniques
- How we will approach each topic
- Reviewing how the Internet works
- Structuring an academic paper
- Delivering a paper presentation

Readings:

1. An Open Platform to Teach How the Internet Practically Works [CCR 2020]
2. How to read a paper
3. LaTeX template
4. Writing tips
5. How to Give a Great Research Talk. YouTube video, Microsoft Research, 2016
6. How to Write a Great Research Paper. YouTube video, Microsoft Research, 2016

Module 2: Surveying the Internet Address Space (I)

In this module, we discuss active measurement scanning techniques across multiple Internet protocols, data collection and analysis to survey the Internet for liveness

(Internet host visibility). We take a look at a proposed taxonomy of liveness that covers responses across multiple protocols. Also, review key findings based on longitudinal studies that cover multiple probe types and replies. Finally, building on the above, we take a look at a technique that predicts Internet services across all ports. The learning framework is first trained on a small sample to learn patterns between services across ports.

Topics:

- Internet scanning as a key Internet measurement technique, and its practical applications.
- Probing techniques to test host responsiveness.
- A method to predict hosted services across ports.
- Good Internet citizenship. Practical and ethical considerations for designing scanning tools and experiments.

Readings:

1. Scanning the Internet for Liveness [ACM CCR, 2018]
2. Predicting IPv4 Services Across All Ports [SIGCOMM, 2022]

Module 3: Surveying the Internet Address Space (II)

In this module, we discuss passive measurement techniques to survey the Internet for "meaningful usage". In other words, we review passive measurements techniques that answer the question "how many of the allocated IPv4 addresses are meaningfully used" and also identify classes of usage, e.g., which address space is used, routed but unused, unrouted and assigned, available? The techniques are based on the analysis of captured network traffic from multiple different vantage points. Finally, we discuss how to detect and deal with traffic that might have forged a source IP address (spoofed), and therefore complicate the data analysis.

Topics:

- A technique for inferring IP address utilization.
- Classifying IP address space by usage.
- A technique to identify spoofed traffic that might be interfering with probing measurements.

Readings:

1. Lost in Space: Improving Inference of IPv4 Address Space Utilization [JSAC, 2016]
2. Detection, Classification, and Analysis of Inter-Domain Traffic with Spoofed Source IP Addresses [Internet Measurement Conference (IMC), 2017]

Module 4: Overtaking the Internet Infrastructure Control

In this module, we take a look at the techniques that are used by cyber actors to overtake (hijack) Internet infrastructure, namely DNS and BGP-based infrastructure hijacking. We take a look at how these techniques work, their flavors and how we can

detect them. Finally, we take a look at a survey that was conducted among network operators about their awareness, the defenses they use in practice and their willingness to adopt new approaches.

Topics:

- How cyberactors can overtake (hijack) Internet infrastructure, namely DNS and BGP-based infrastructure.
- Example approaches to detect hijacking.

Readings:

1. Retroactive Identification of Targeted DNS Infrastructure Hijacking [Internet Measurement Conference (IMC), 2022]
2. Profiling BGP Serial Hijackers: Capturing Persistent Misbehavior in the Global Routing Table [Internet Measurement Conference (IMC), 2019]

Module 5: Internet Censorship I

In this module, we discuss Internet censorship by reviewing the blocking methods and censoring devices that are used by censors worldwide. We review measurement techniques and datasets to understand emerging censorship techniques: 1) Geoblocking, where users from particular countries or regions are denied access to particular servers, and 2) Twitter throttling, where users experience significant slowdown as a means of discouragement. Finally, we look into measurement approaches that are targeting to locate censoring devices and their operations across multiple countries and networks.

Topics:

- What is Internet censorship: An overview of content blocking methods and approaches.
- A technique to identify and locate censorship devices

Readings:

1. Network Measurement Methods for Locating and Examining Censorship Devices [CONEXT, 2022]

Module 6: Internet Censorship II

In this module, we discuss an Internet measurement platform that is specialized for censorship research, namely the Censored Planet. We take a look at the architecture, ongoing deployments, data collection and analysis techniques that are used to study censorship over multiple geographic regions and networks, as well as the resulting longitudinal studies.

Topics:

- What are Internet censorship observatories? An example observatory and how it is designed.

- A technique to detect censorship at scale. From rule-based to ML-based censorship detection.

Readings:

- Censored Planet: An Internet-wide, Longitudinal Censorship Observatory [CCS, 2020]
- Augmenting Rule-based DNS Censorship Detection at Scale with Machine Learning [ACM SGKDD, 2023]

Module 7: Web and Trust Management

In this module, we discuss the certificate authority ecosystem by looking into measurement techniques that are based on the collection and analysis of multiple datasets: certificate snapshots, IPv4 scans, popular domains (e.g., Alexa), certificate logs, and others. We also discuss how previously issued certificates are revoked as well as the associated risks of stale certificates.

Topics:

- How certificate issuance works.
- A measurement approach to identify certificates that are stale or revoked certificates, and how they can be abused.

Readings:

1. Stale TLS Certificates: Investigating Precarious Third-Party Access to Valid TLS Keys [IMC, 2023]
2. Towards a complete view of the certificate ecosystem. [IMC, 2016]
3. An End-to-End Measurement of Certificate Revocation in the Web's PKI [IMC, 2015]

Module 8: Measurements & Voting Systems

In this module, we review online voting systems and their associated risks. Making remote voter participation easier and available to more people is an advantage. But at the same time online ballot delivery introduces new risks that could even alter major election results. We review the Omniballot platform and the associated risks from the client side.

Topics:

- How certificate issuance works.
- A measurement approach to identify certificates that are stale or revoked certificates, and how they can be abused.

Readings:

1. Security Analysis of the Democracy Live Online Voting System [Initial public version] [USENIX, 2021]
2. Can Voters Detect Malicious Manipulation of Ballot Marking Devices? [OAKLAND, 2020]

Module 9: The Landscape of Abuse on Social Platforms

In this module, we review the landscape of abuse and threats on social platforms. We focus on toxic content and discuss the behaviors of online toxic accounts that are related to hate and harassment, and how they may be related to other emerging online threats. We discuss how to study the behaviors of these accounts by looking at an example study on Reddit toxic accounts; from data collection to studying trends and patterns. We also discuss how to approach designing a toxic content classifier and the challenges to consider.

Topics:

- An overview of abuse on social platforms with a focus on toxicity, hate and harassment.
- A method to study behaviors, trends and patterns of toxic accounts.
- Example strategies to identify toxic content.

Readings:

1. SoK: Hate, Harassment, and the changing landscape of Online Abuse. [OAKLAND, 2021]
2. Understanding the Behaviors of Toxic Accounts on Reddit. [WWW, 2023]
3. Designing Toxic Content Classification for a Diversity of Perspectives [USENIX SOUPS, 2021]

Module 10: Entities on Social Platforms

In this module, we discuss abusive accounts on social platforms that are related to a wide spectrum of abuse (or violations) such as spam, fake engagement, illegal content, violence, and terrorism. We start with a taxonomy of these accounts and the challenges to detecting them at scale, and we provide an overview of existing detection approaches. Then, we focus on holistic approaches that leverage the network structure of the social networks on which accounts operate. These approaches suggest capturing the overall behavior of the account of interest by creating embeddings of the account, taking into consideration the network structure, properties, and behaviors of the neighboring accounts as well. We also discuss how to design a system based on this approach, how to extract features, and how to perform evaluation. Finally, we consider a study that leverages the social network structure to detect multiple identities in discussion communities.

Topics:

- A taxonomy of abusive accounts and the spectrum of their behaviors.
- A technique to study sock puppet accounts
- An example technique to detect abusive accounts at scale.

Readings:

1. An Army of Me: Sockpuppets in Online Discussion Communities. [WWW, 2017]

2. Deep Entity Classification: Abusive Account Detection for Online Social Networks [USENIX, 2021]

Module 11: False Information on the Web and Social Platforms

In this module, we review the landscape of false information on the web and social platforms, including the mechanisms, actors, rationale and characteristics. We also review detection approaches. Further, we study how to identify false information campaigns that spread across multiple platforms (including cross-references in text and videos). We also focus on an example study that presents data collection, feature extraction and system evaluation.

Topics:

- An overview of false information on the web and social platforms.
- An example technique to detect cross-platform spread.

Readings:

1. False Information on Web and Social Media: A Survey [Book chapter, CRC Press, 2018]
2. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos [ICWSM, 2022]

Module 12: The False Information Ecosystem

In this module, we discuss the ecosystem that supports false information. More specifically, we first discuss the web hosting ecosystem that supports false information, such as websites related to narratives and conspiracy theories. We also discuss a web crawling technique to collect data, how to form a graph based on the collected data, and perform analysis on the graph. Next, we look at the Internet infrastructure that supports false information websites, including domain registrars, web hosting and email providers. We will talk about an example study that shows how measurements can be used to identify and study the underlying Internet infrastructure.

Topics:

- A technique to identify Internet infrastructure (domain registrars, email providers, advertising partners) that supports websites with false information (e.g. websites with narratives, conspiracy theories).
- A graph-based technique to understand the underlying relationship between news websites and false information websites.

Readings:

1. On the Infrastructure Providers that Support Misinformation Websites [ICWSM, 2022]
2. No Calm in The Storm: Investigating QAnon Website Relationships [ICWSM, 2022]

Module 13: Online Social Platforms as a Vantage Point to Study Online Cybecrime Communities

In this module, we discuss the opportunities to leverage online social platforms as vantage points to detect emerging threats. More specifically, we take a closer look into two types of platforms: 1) GitHub repositories, and 2) channels of online chatter, e.g., tweets, blogs, and underground forums. We will examine how to mine each platform with the goal of identifying emerging threats, encompassing data collection, preprocessing, feature engineering, designing a learning system, and evaluating the proposed framework.

Topics:

- Online cybercrime communities: how different disciplines approach to understanding them.
- Mining online platforms.
- An example technique to mine the GitHub platform to identify potentially suspicious repositories.

Readings:

1. SourceFinder: Finding Malware Source-Code from Publicly Available Repositories [USENIX RAID, 2020]
2. The Art of Cybercrime Community Research [ACM Computing Surveys, 2024]

Module 14: Ethics in Internet Measurements

In this module, we discuss ethical challenges and risks when collecting, analyzing, sharing, or publishing network data. Myriads of research projects, which have been advancing our understanding of an ever-expanding list of topics, have been performed based on the collection and analysis of sensitive network data. We review the ethics issues and discuss the ethics principles that include: informed consent, human rights, releasing and using shared data, hacking, analysis techniques, ethical review, and Research Ethics Boards (REBs). We talk about two case studies (malware exploitation, classified materials) that highlight the associated challenges and risks in practice.

Topics:

- Overview of ethical challenges and risks faced by measurement studies, including sensitive Internet data
- Overview of ethics norms: informed consent, human rights, releasing and using shared data, hacking, analysis techniques, ethical review, and Research Ethics Boards (REBs)

Readings:

1. Ethical issues in research using datasets of illicit origin [IMC, 2017]
2. Understanding the Ethical Frameworks of Internet Measurement Studies [NDSS, 2023]

Module 15: Sustainable Research: The Importance of Transparency, Reproducibility & Replicability

In this module, we discuss the importance of reproducibility in scientific research since it is crucial to accelerate the transition between science and technology and establish the trustworthiness of results. In this context, we review how ACM defines the cornerstones of sustainable research, namely: repeatability, reproducibility, replicability, their goals and their principles. We also discuss the challenges towards reproducibility. We focus on the best practices as recommended by academics and practitioners. Finally, we discuss the importance of good documentation and other practices through a case study of a large-scale web measurement.

Topics:

- Goals & challenges of reproducible research
- ACM definitions: repeatability, reproducibility, replicability
- Best practices for scientific Internet research

Readings:

- The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research [ACM CCR, 2019]
- Encouraging Reproducibility in Scientific Research of the Internet [NSF Seminar, 2018]
- Reproducibility and Replicability of Web Measurement Studies [WWW, 2022]

Course Policies

Late submissions & extensions. The students are expected to complete the work on time by the due dates. In case of an emergency, please reach out to TA team through a private Ed Stem post, so we can come up with a plan to make up for the work or alternative solutions, depending on the type of the emergency and the impact it has.

Plagiarism & academic integrity. Students are expected to follow the Georgia Tech Honor Code (https://policylibrary.gatech.edu/student-life/academic-honor-code), including the Graduate Addendum. All incidents of suspected dishonesty will be reported to and handled by the Office of Student Integrity. If in doubt to whether an action is allowed in this course, please ask the Instructor/TAs.

In addition, the following specific policies apply to this course:

1. If your MIRM project is the same or similar to a project that you are working on in the current semester or you have worked on in previous semesters (e.g. other class projects, or 8903s, etc.), you must communicate with the instructional team to carve out a piece that is appropriate and specific for the MIRM course for the current semester.
2. ACM guidelines on Authorship and Acknowledgements: https://www.acm.org/publications/policies/new-acm-policy-on-authorship
3. ACM guideline on the use of AI: https://www.acm.org/publications/policies/frequently-asked-questions
4. According to the ACM guidelines for Acknowledgements, we ask that you acknowledge the MIRM course for providing guidance for your paper. For example, the following statement is sufficient to acknowledge our course's contribution in guidance.

   "We would like to thank the Georgia Tech OMSCS-8803-O23 Modern Internet Research Methods course instructional team for supervising the first draft of this paper."

5. To help inform prospective students about the types of papers the MIRM instructional team works on with the students, we plan to include a high-level description of each semester's projects on the course's website.

Ed Discussion code of conduct. The students are expected to be respectful to others when interacting on Ed Discussion. Please review the students' code of conduct. https://policylibrary.gatech.edu/student-life/student-code-conduct

Georgia Tech student resources & student accommodation. The Disability Services team collaborates with the students to find creative solutions and reasonable accommodation. Please contact the Office of Disability Services at (404)894-2563 or http://disabilityservices.gatech.edu/, as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodation letter. Please also send a private message to "Instructors" on Ed Stem as soon as possible. Please note that the

TA team is not able to provide any accommodation or extensions without an accommodation letter, nor the accommodations can be provided retroactively.

Communication Policy

Please use Ed Discussion (available via Canvas course site) for all communication with the instructional team.

Subject to Change Note

Please note that the current syllabus is subject to change at any time.

Contributions

See Extra Credit III assignment.