CS 8803 Internet Data Science

Alberto Dainotti
1/10/2022

Georgia Tech

Welcome!

Overview of today

- COVID, Safety, Health
- People Introduction
- Course goals
- Course logistics
- Course format / assignments / grading
- Action points

Thursday

- Lecture on the course topics

COVID & our safety

- Attendance
  - In-person attendance is not required. You may join online
  - Attendance either in-person or online IS required
  - This is a discussion-oriented course
    - To encourage participation, meeting **recordings WILL NOT be regularly shared**. I will do my best to share recordings on demand only to students who have a (documented + good) reason for not being able to attend on that specific day. We will be both **synchronous in online and in person** for both Tuesday and Thursday lectures. Please note that participation is part of the course grade, so those joining remotely are expected to participate in in-class discussion.
- Safety measures
  - Masks
    - Public health evidence overwhelmingly indicates that masks are key to reducing transmission and incidence of COVID. If you attend class in person, I strongly request that you wear a mask to keep yourself and others as safe as we can. I will be wearing a mask at all times.

COVID & our safety

- Alternative classroom?
    - ES&T L1116

Ford Environmental Science & Technology

311 Ferst Drive

COVID & our safety

- Safety measures [continued]
  - To enable online participation we need to use microphones. Besides the presenter's microphone there will be a **shared microphone** in the classroom *(unless ceiling microphones are installed)*.
  - CDC:
    - "The principal mode by which people are infected with SARS-CoV-2 (the virus that causes COVID-19) is through exposure to respiratory droplets carrying infectious virus."
    - "It is possible for people to be infected through contact with contaminated surfaces or objects (*fomites*), but the risk is generally considered to be low."
      - "Case reports indicate that SARS-CoV-2 is transmitted between people by touching surfaces an ill person has recently coughed or sneezed on, and then directly touching the mouth, nose, or eyes. Hand hygiene is a barrier to fomite transmission and has been associated with lower risk of infection."
    - *https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/surface-transmission.html*
  - You're encouraged to use hand sanitizers when entering the classroom, before touching a microphone, …

COVID & our safety

- Having questions/concerns is normal - *https://care.gatech.edu*

Welcome!

- You are brave!
  - This class is an experiment in itself
- Ever heard about “Internet Data Science” ?
  - Me neither :-)
- First time a graduate research class is called this way. First time teaching it.
  - An opportunity to work with some new concepts and tools or to see them in a different light.
  - Slightly different than classic paper-only format
  - Collaborative effort

Internet Data Science?

- Combination of
  - Networking -> Internet **Measurement** and **Data**
  - Data Science -> **Tools/Methods**
  - Other disciplines -> **Questions**/methods/**data**
  - Interdisciplinary questions:
    - *Society: digital divide, human rights, privacy, democracy, bias, …*
    - *Law: public policy, cybercrime, …*
    - *Other engineering fields: reliability of power grids, …*
    - *International relations: cyberwarfare & deterrence, influence, …*
  - *Also more technical/CS questions related to the Internet utilization and infrastructure*

Who am I?

- Alberto Dainotti
- Associate Professor, School of Computer Science
- Born and raised in Italy
  - PhD University of Napoli, Federico II, Italy
- Last 10 years at Univ. California, San Diego
- Just moved to GATech!
  - Started the Internet Intelligence Lab @ SCS
  - Research focus: understand and improve Internet reliability and security through measurements and data analysis/analytics; Internet infrastructural characteristics and evolution; Interdisciplinary work

Dr. Zachary Bischof

- Research Faculty, School of Computer Science
- PhD Northwestern University
- Remotely working from Oregon
- Previously a post-doctoral fellow at IIJ research lab in Tokyo, JP
- Just joined GATech as well!
  - and joined the Internet Intelligence Lab @ SCS

Who are you?

- Your name*
- Your program
  - Which year/When do you expect to graduate
- Have you taken a research seminar class before?
- Have you ever written a research paper?
- Random interesting fact about you (if you choose to share)
- What do you hope to get out of this class

* *Forgive me for mispronunciations or poor memory!*

Course goals

- Understand the *dimensions* along which Internet-related phenomena can be analyzed and the related datasets
  - Their characteristics, complexities/limitations/caveats, …
  - What data is available
  - How more data can be generated or obtained
- Develop skills to critically evaluate empirical methods and analyses
- Identify interesting research questions and topics
- Execute a quality research project (e.g., think workshop paper)
- Have a fun semester learning about interesting topics
- Contribute to define this scientific discipline / shape this class

Course logistics / requisites etc.

- Class meetings
  - When: T/Th 5:00—6:15pm
  - Where: Van Leer E361
- Office Hours: Wed 3-4pm EST KACB-3336/Zoom (or by appointment)
    - Week 1: Fri 14th 3-4pm EST KACB-3336/Zoom
- Course resources
  - Piazza, accessible within **Canvas** (discussions/questions)
  - Canvas (all announcements, assignments, grades, surveys)
  - Course Website, linked to within Canvas (official syllabus/schedule)
    - *https://www.cc.gatech.edu/~adainotti6/cs8803-ids*
  - Course Notion (website->schedule/syllabus->Notion page w/ schedule -> subpages)
- Email me directly: dainotti@gatech.edu

Notion (www.notion.so)

- The schedule with the list of papers is in Notion and public
- Additional pages pointed from there whose access is restricted
- I will grant you “edit" access on the restricted-access pages

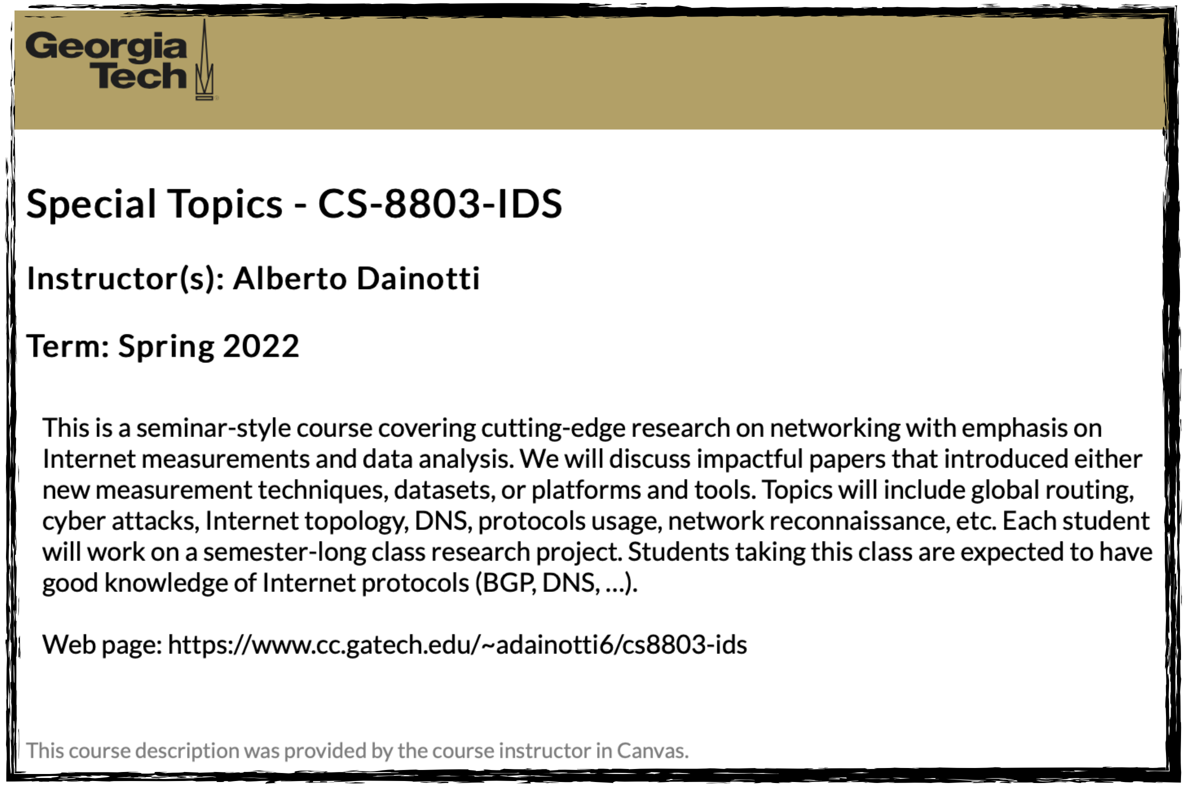
- Hoping next year to migrate to something more “open”: e.g., *anytype.io*

Prerequisites

- Prerequisite: Undergraduate Networking
  - CS 3251 Undergraduate Computer Networking or equivalent

- Helpful:
  - CS 4251/6250 Computer Networking II / Graduate-level Computer Networks

- Background:
  - IP subnetting, IPv6, DNS, BGP, Autonomous Systems, TCP/UDP/ICMP, traceroute, …



Georgia Tech

Special Topics - CS-8803-IDS

Instructor(s): Alberto Dainotti

Term: Spring 2022

This is a seminar-style course covering cutting-edge research on networking with emphasis on Internet measurements and data analysis. We will discuss impactful papers that introduced either new measurement techniques, datasets, or platforms and tools. Topics will include global routing, cyber attacks, Internet topology, DNS, protocols usage, network reconnaissance, etc. Each student will work on a semester-long class research project. Students taking this class are expected to have good knowledge of Internet protocols (BGP, DNS, ...).

Web page: https://www.cc.gatech.edu/~adainotti6/cs8803-ids

This course description was provided by the course instructor in Canvas.

Format *[1/2]*

- Nothing like today or this Thursday!
- Course driven by **discussion of** academic **papers + datasets**
- Homework:
  - (Paper readings + Checking a dataset) + analytical summary write-ups on those papers/datasets, driven by standard questions
- Class:
  - 1-2 students prepare and present a 25 minute presentation summarizing the class's reading, and help lead the class discussion.
    - Can be longer if you need significant extra time to introduce background concepts
- We read papers, check/try datasets, talk about them critically. I help you guide the discussion

Format *[2/2]*

- You eventually apply what you've learned as a project
  - Write and present a project proposal (due 3/1)
  - Present a talk and/or demo on your final project (due 4/19)
  - Submit a research-style paper/notebook on your final project (due 4/29)

- The project/paper does *not need* to be like the papers you read
  - It needs to leverage the types of data we discuss and the knowledge you acquire during the class
  - It can be a DeepNote/Colab notebook
    - E.g., take a look at *https://github.com/britram/trilateration/blob/master/paper.ipynb*
  - It can be a new dataset, or a careful validation/update of an existing dataset, or a novel interesting visualization
  - …

Internet Data Science

Paper readings

- Combination of
  - Networking -> Internet **Measurement** **ata**
  - Data Science -> **Tools/Methods**
  - Other disciplines -> **Questions/** **/data**
    - *Society: digital divide, human rights, privacy, democracy, bias, …*
    - *Law: public poli ime*
    - *Other engin grids, …*
    - *Internation nce, influence, …*
    - *Also more t nternet utilization and infrastr*

Paper readings

Projects

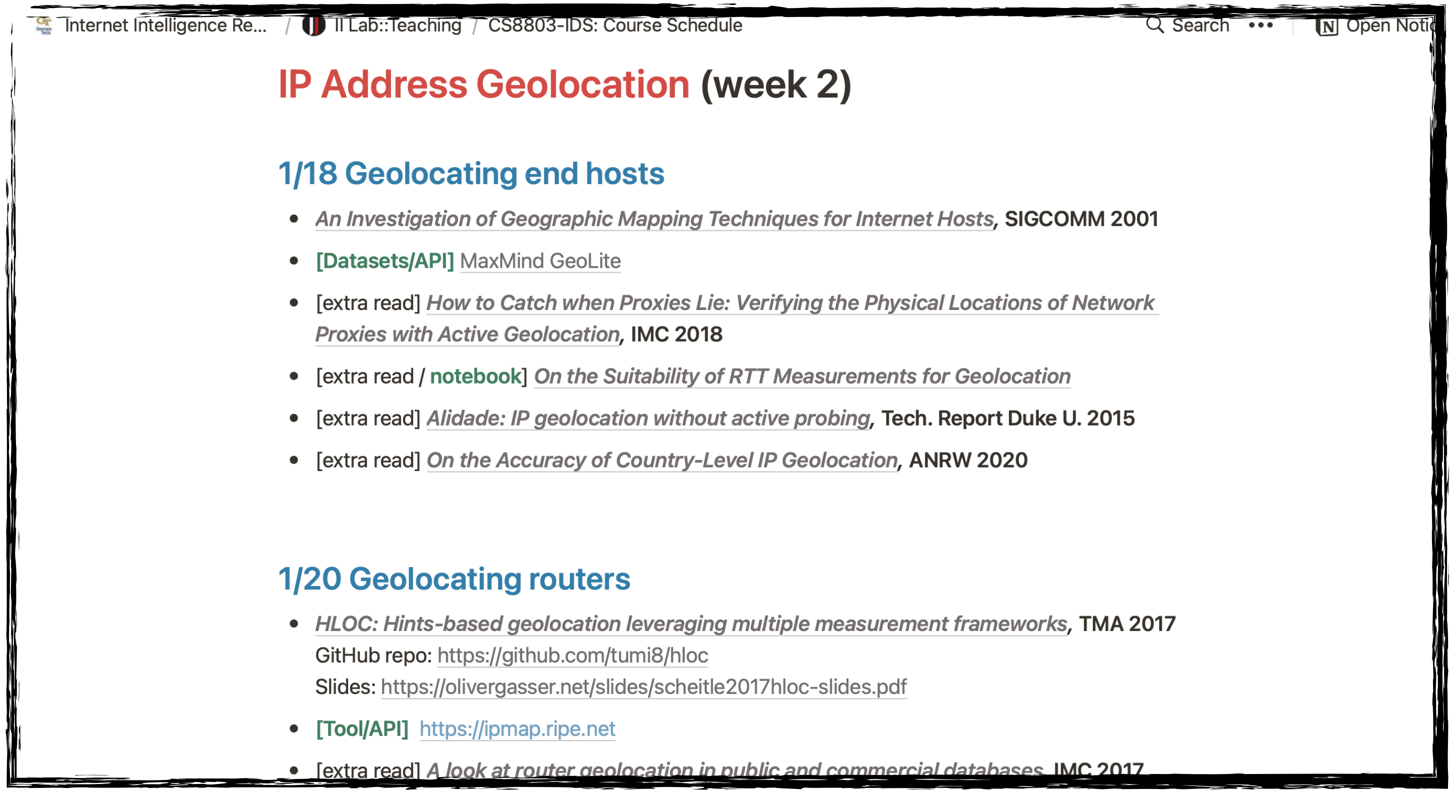
Projects

Class components and weight

- Participation (**30%**)
  - Attend and engage with class meetings (ask and answer questions, provide comments).
  - Posting and answering questions on Piazza
  - Contributing (in the Notion pages) with links, code samples/notebooks, and useful commentary on datasets
- Discussion Lead (**20%**)
  - For one class during the semester, prepare and present a ~25 minute presentation summarizing the class's reading, and help lead the class discussion. **Two class presentations if you co-present.**
- Paper and Dataset Summaries (written) (**20%**)
  - For each assigned paper and dataset, submit a brief paper analysis based on a set of standard questions provided by the instructor. **In total, 2 summaries can be skipped during the semester (you still need to email the instructor on time stating you skip).**
- Final Project (**30%**)
  - Project Proposal (10%) - Write and present a project proposal (due 3/1).
  - Project Presentation (10%) - Present a talk and/or demo on your final project (due 4/19)
  - Project Writeup (10%) - Submit a research-style paper on your final project (due 4/29)

Current papers/datasets schedule

- The list is still fluid
  - The assignments will normally be 1 paper + 1 dataset
  - We'll add more datasets as pointers to concepts and to give you ideas about projects (no extra homework!)


Internet Intelligence Re... / II Lab::Teaching / CS8803-IDS: Course Schedule
Search
•••
Open Noti

IP Address Geolocation (week 2)

1/18 Geolocating end hosts

- *An Investigation of Geographic Mapping Techniques for Internet Hosts*, SIGCOMM 2001
- [Datasets/API] MaxMind GeoLite
- [extra read] *How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation*, IMC 2018
- [extra read / notebook] *On the Suitability of RTT Measurements for Geolocation*
- [extra read] *Alidade: IP geolocation without active probing*, Tech. Report Duke U. 2015
- [extra read] *On the Accuracy of Country-Level IP Geolocation*, ANRW 2020

1/20 Geolocating routers

- *HLOC: Hints-based geolocation leveraging multiple measurement frameworks*, TMA 2017
  GitHub repo: https://github.com/tumi8/hloc
  Slides: https://olivergasser.net/slides/scheitle2017hloc-slides.pdf
- [Tool/API] https://ipmap.ripe.net
- [extra read] *A look at router geolocation in public and commercial databases*, IMC 2017

Extensions and Late Assignments

Assignments are due at the time listed in the schedule. There are no undocumented exceptions. If you have an emergency situation or a school sanctioned exception, please contact me before the due date, so we can adjust your assignment deadlines (some documentation may be needed).

In total, 2 summaries/analyses can be skipped during the semester (you still need to email the instructor on time stating you skip).

Papers & datasets/tools/APIs summaries *[1/2]*

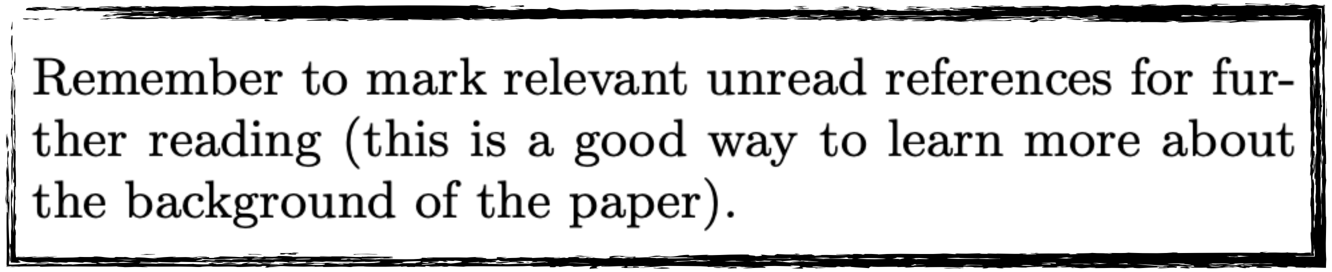
- Papers: **Thoughtful** (but concise, 2-5 sentences per answer) reflection on the paper reading.
  - What are the paper's contributions?
  - What did you like about the paper?
  - What are questionable parts of the paper and its limitations? (E.g., methodology issues, detail omissions, presentation problems)
  - What was unclear about the work, or what questions do you have?
  - List the key datasets used/obtained in the paper.
    - Extra points: list here and add to Notion useful+accessible datasets related to the topic
  - Some readings might have a specific question or two.
- Due **5pm the day before class**. Submit to Canvas assignment *(but work on it elsewhere)*

Papers & datasets/tools/APIs summaries *[2/2]*

- Datasets/Tools/APIs:
  - Were you able to download the dataset / try the tool? *[be determined*]*
  - What was the simplest thing you were able to do with it?
  - Limitations & strengths:
    - Accessibility of the data (license terms, immediate access vs registration, free,). Can it be used/accessed from a Notebook? (e.g., with a subscription key, through an API, simply downloading the dataset, …)
    - Is the methodology fully explained and sound?
    - Is the data (or the data availability) limited to certain time frames? E.g., only recent snapshots and no historical data, or only old data, …
    - Strengths or ideas of how you could use the dataset
  - Extra points: Have you found alternative/similar/complementary datasets? (List them + add them to the Notion page)
- Due **5pm the day before class**. Submit to Canvas assignment *(but work on it elsewhere)*

Reading papers

- Reference conferences / workshops
- *"How to Read a Paper", S. Keshav, http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf*



Remember to mark relevant unread references for further reading (this is a good way to learn more about the background of the paper).

- Extra: to get a sense of the progress of the state of the art, use google scholar to find citations of the paper you're reading and see if you find interesting papers that came after that

Discussion Leads

- Discussion Lead: Prepare a ~25 minute presentation discussing the paper. Assume folks have read the paper (but don't remember every detail).
  - 1-2 students lead per class. Use slides.
- Don't need to be terribly pretty, but effective communication is important!! You are allowed to use the authors' slides.
- Email me your slides the night before class
- Rest of the class (including me): Ask questions and make comments.
- Key parts: Background, Problem/Motivation, Relationship to Prior Work, Methodology, Evaluation/Results, Implications, Paper Summary Questions, Datasets

Participation

- Expectations:
  - You attend class regularly
  - You have read the paper
  - You have answered the question(s)
  - You constructively participate in discussions
- I normally don't cold-call anyone, it's up to you to join in
- Good:
  - “I didn't understand X”
  - “I thought Y was neat”
- Bad:
  - “This author is stupid”
  - “This work is pointless”
  - I never see you again after today but you appear on the roster in April

Project timeline

- Group size: 2 people (group of 1 is ok but needs a good reason)

- Project **Proposal** - Write and present a project proposal (due 3/1)
- Project **Presentation** - Present a talk and/or demo on your final project (due 4/19)
- Project **Writeup** - Submit a research-style paper/notebook on your final project (due 4/29)
- Details will be posted as Canvas assignments

- *Note: You should be engaging with me throughout the semester for feedback/direction*

Tips

- Ask questions
- Come to my office hours
- Engage critically with the readings and discussions
  - This is an acquired skill. The paper summaries and class discussion are to help you.
- Don't procrastinate on the projects!
  - Get feedback early, get it often!
- Give me feedback!
  - I want you to learn and do well. I want this class to be better. I'll ask for anonymous feedback

Action points for you

- Read the website and check again these slides
- Join Canvas + Piazza

- Look over the class topics and think of the ones you might be interested in leading. I will send a schedule request soon.
- Don't read the paper for next Tuesday yet. We'll finalize Tuesday's paper this Thursday

- Ask me ASAP if anything is unclear or missing!
  - Or if you have doubts about this class being for you

Thanks

Credits: slides text on course organization partially based on text/ideas from Frank Li & Paul Pearce