Metadata and Content

What is Metadata? What is Content?

- It often matters from a legal perspective
- It’s often important for privacy
- Most people are unaware of metadata
- We’ll discuss three types: map apps, photos, Internet communications

Mapping Apps

- Most are client-server: you enter data on your phone, but a provider—Google, Apple, Microsoft, etc.—supplies the maps, calculates the routes, warns you of traffic, etc.
- What do you know you're sending the server?
- What does it actually know?

Obviously Supplied

- Starting point
  - Often, that's the current location
- Destination
- Mode of transportation
- And: current location as you travel
- (Maybe other data from your phone, but that's a privacy issue, not metadata…)

What Does the Server Know?

- Speed? Absolutely
  - Map servers use that to indicate traffic jams and estimate route time
- Common routes used? Probably—use that to fine-tune algorithms
- Acceleration? Phones do have accelerometers
  - Is your insurance company interested in how rapidly you accelerate and how hard you brake?
- Time of day?

Photo Metadata

What else is there besides the picture?

Bighorn sheep, Rocky Mountain National Park, June 3, 2019

EXIF Metadata

- EXIF: Exchangeable Image Format
- Metadata standard for many image and audio file formats
- Contains a vast amount of data

Big horn sheep, Rocky Mountain National Park, June 2, 2019

General | Exif | GPS | IPTC | JFIF | TIFF

Aperture Value	**4.971**
Body Serial Number	**2616090**
Color Space	**sRGB**
Components Configuration	**1, 2, 3, 0**
Contrast	**Normal**
Custom Rendered	**Normal process**
Date Time Digitized	**Jun 3, 2019 at 8:46:18...**
Date Time Original	**Jun 3, 2019 at 8:46:18...**
Digital Zoom Ratio	**1**
Exif Version	**2.3.1**
Exposure Bias Value	**0**
Exposure Mode	**Auto exposure**
Exposure Program	**Aperture priority**
Exposure Time	**1/8000**
File Source	**DSC**
Flash	**Off, did not fire**
FlashPix Version	**1.0**
FNumber	**5.6**
Focal Length	**300**
Focal Length In 35mm Film	**450**
Focal Plane Resolution Unit	**centimeters**
Focal Plane X Resolution	**2,558.641**
Focal Plane Y Resolution	**2,558.641**
Gain Control	**Low gain up**
Lens Model	**70.0-300.0 mm f/4.5-5.6**
Lens Specification	**70, 300, 4.5, 5.6**
Light Source	**unknown**
Max Aperture Value	**5**
Metering Mode	**Spot**

EXIF Data

- Standard photographic information: shutter speed, aperture, lens focal length
- Less common photographic information: “Y Cb Cr Sub Sampling”, “Green Matrix Column”
- >300 data items, according to exiftool
- Privacy-sensitive items: body and lens serial numbers

Location!

- Precise GPS location
  - Most phones will add it by default—but I add it manually to my nature pictures…
- Sensitive enough that Macs, at least, make it very easy to strip

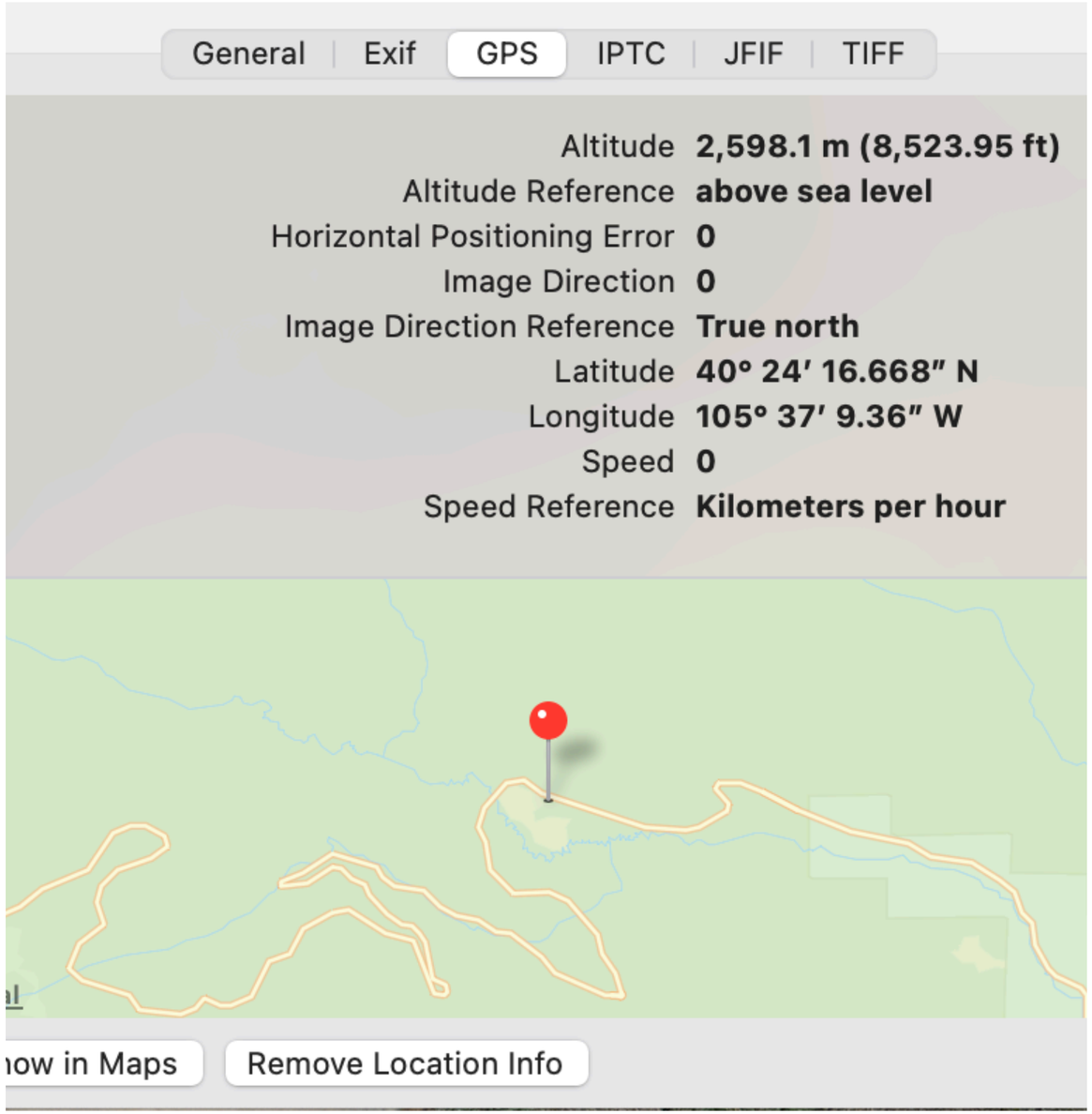


Photo Metadata

- When you share a photo, you share its metadata
- Twitter and Facebook strip all metadata—but what do they do with it?
- When law enforcement seizes photos from a device, the metadata goes along with the pictures

What is Metadata, *Legally*?

- To me: data that you don't know that you're sharing
- Sometimes, *knowing* conveyance is critical, per *Smith*
- But there isn't a bright line!

Content and Metadata in the Phone Network

- In *Katz* (1967), the Supreme Court overruled *Olmstead* (1928) and held that the contents of phone calls were protected by the Fourth Amendment
- In *Smith* (1979), the Court held that dialed numbers were not similarly protected, due to the third party doctrine
- The entire model of content versus metadata protection is based on 40+ year-old telephony
- Does the model work for the Internet?
- (As of about 1979, it didn't even work properly for the phone network…)

From *Smith*

"All telephone users realize that they must 'convey' phone numbers to the telephone company, since it is through telephone company switching equipment that their calls are completed. All subscribers realize, moreover, that the phone company has facilities for making permanent records of the numbers they dial, for they see a list of their long-distance (toll) calls on their monthly bills."

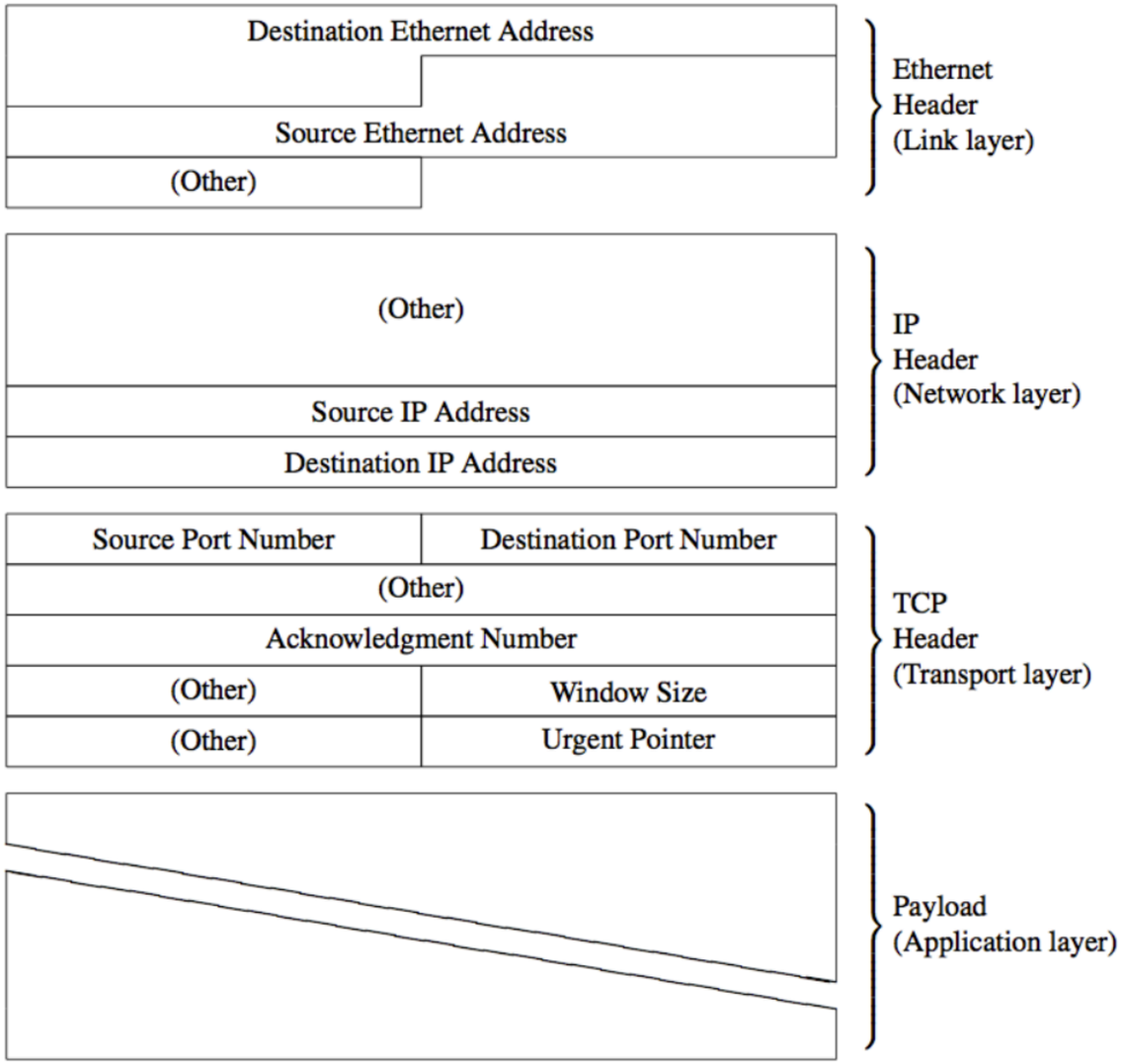
"This Court consistently has held that a person has no legitimate expectation of privacy in information he voluntarily turns over to third parties."

On the Internet, what information do people "realize" that they "voluntarily" send to a third party?

CS @CU

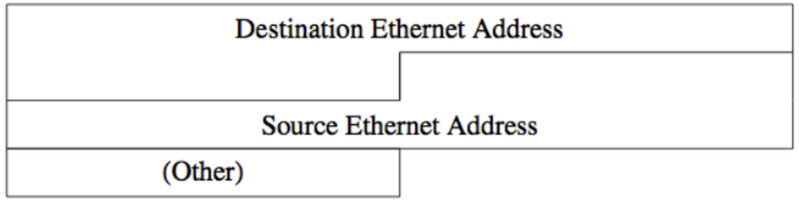
Addresses

- (Almost) every packet has (at least) three sets of addresses: Ethernet (MAC), IP, and port numbers
- To whom are these addresses given?
- Which are knowingly and voluntarily given?



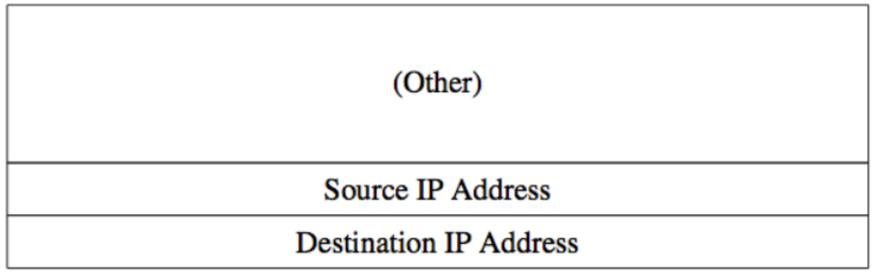
Ethernet (MAC) Addresses

- MAC addresses stay on-network, e.g., your home WiFi network or a public hotspot
- If you rent a modem from your ISP, does your ISP know your MAC address?
  - From a technical perspective, it *must* be sent *to* the modem, but is it forwarded and stored by your ISP?
- How many people even know what a MAC address is?
  - Btw: your phone sends a MAC address when on WiFi, but not when using cellular data—but for cellular data, there are other identifying numbers



IP Addresses

- IP addresses are in every packet
- They're seen and used by every ISP along the path from you to your destination
- They're a clear analog to phone numbers—but have you ever seen a list of destination IP addresses on your monthly bill? Do you know if your ISP has the technical ability to collect them?
- By the way, what is your laptop's or phone's IP address? It's not easy to find out, and for technical reasons ("Network Address Translators"—NATs) what you see may not be what the other end of a connection sees



TCP Port Numbers

- Port numbers are in most packets
- If an IP address is like a street address, a port number is like the room within that building
  - “Room” 25 is the mailroom, “room” 80 (HTTP) is the library, etc.
- Again, most people don’t know what port numbers are—but it’s more complicated than that, and the problems are deeper

Source Port Number	Destination Port Number
(Other)	
Acknowledgment Number	
(Other)	Window Size
(Other)	Urgent Pointer

TCP Port Numbers

- From the strict, 1981 definition of TCP, port numbers are “end-to-end”—they’re of no concern to intermediate routers along the path, and aren’t used by them
- In other words, they’re not third-party data—or are they?
- Every commercial router ever built has the capability of looking at, and filtering on, TCP port numbers
- ISPs often monitor customers’ port numbers to understand traffic patterns
- Under certain circumstances, ISPs have to look at (but not permanently record) port numbers to help customers’ TCP performance

And it’s worse than that…

TCP Port Numbers, Hotspots, and Cellular

- Because the world has run out of IP addresses, IP addresses are generally translated at the originating network's border
  - Doing this translation *requires* use of packets' port numbers
- This happens with *all* cellular data connections
- It happens at *all* public hotspots
- It happens with your home network—and your modem generally does the translation
- At Columbia, it does *not* happen on the campus WiFi network—but does happen on the Law School's network
- On the bridge over Amsterdam Avenue, how do your packets flow? CUIT? Law? Phone?

TCP Port Numbers

- TCP port numbers are essentially never *given* to third parties
- They're frequently *taken* by such parties
- It takes a great deal of sophistication to detect NATs
- It's impossible to detect other use or collection of port numbers
- Are they third-party data? DoJ says yes, the Internet architecture says no, but the real-world technology is very complex, and there's no case law
- (And again, you don't see port numbers on your monthly bill…)

Email

- Is email protected by the Fourth Amendment?
  - Is email “persons, houses, papers, and effects”?
- Search of email less than 180 days old requires a warrant, per statute (18 U.S.C. §2703(a))—but older email may require just a subpoena (§2703(b))
- The Sixth Circuit held that all email is protected by the Fourth Amendment (United States v. Warshak, 631 F.3d 266, (CA6, 2010))
- Does your mail provider do spam filtering? Virus-scanning? Does that mean you’ve given it to the provider, and hence have lost all privacy interest in it?

From *Smith v. Maryland*

police the numbers he dialed. The switching equipment that processed those numbers is merely the modern counterpart of the operator who, in an earlier day, personally completed calls for the subscriber. Petitioner concedes that if he had placed his calls through an operator, he could claim no legitimate expectation of privacy. Tr. of Oral Arg. 3-5, 11-12, 32. We are not inclined to hold that a different constitutional result is required because the telephone company has decided to automate.

But what about spam filtering for email?
Does that make a difference? Should it?

Envelopes and Contents

- When you send email, the protocol has an “envelope” portion, listing the addresses, and “contents”, the message itself including the familiar header lines (`From:`, `To:`, etc.)
- The envelope also contains `From:` and `To:` lines
- If a mail service such as `gmail` is used, the envelope lines are clearly third party data
- What about the corresponding lines in the contents? Obviously not—but DoJ claims otherwise

More Email Third Party Issues

- I run my own mail server
- A friend (and frequent co-author) runs one, too
- On a recent paper, our co-authors all had email addresses on my mail server
- If my friend sent email to all of us, there would be no third party on a message to me, but there would be on messages to our co-authors
- You can't tell if it's third-party data until after you've intercepted it…

Intuition to the Rescue?

- Suppose, when you make a phone call, you spoke a different language? Would it work? Of course.
- Suppose that when dialing a call, you sent different tones for the digits. Would that work? No, of course not
- Can we use this—sensitivity to the “language” spoken—to distinguish content from metadata?
  - (Instead of “language”, could we use encryption?)
- Sometimes that works—but not always

Spoken Language and Metadata

- Modern voice communications—mobile phones and voice over IP (VoIP)—compress the sounds
- The actual compression, e.g., the size and timing of the compressed packets, depends on the sounds
- These sounds are different for different languages!
- Metadata thus reveals which language is being spoken and sometimes even phrases

Email Headers

- Email contents consist of a header component and a body component
- All of it is end-to-end, and clearly not third-party data
- However, the header contains sections that are examined and modified by mail servers
- If the header isn't formatted correctly—or is encrypted or is in another "language"—the mail server can't cope
- In other words, *part* of the header might be considered third-party content

The Web

URLs: https://www.example.com/more/path?query

- IP addresses are third party data, but many hostnames can have the same IP address
- When you click on something, you don't know if the `GET` method or the `POST` method is used—with the former, data entered on forms becomes part of the next URL, and hence visible to others (and possibly be third party data)
  - But this the web site designer's choice!
- There are sometimes alternate forms, e.g., `patents.google.com` versus `google.com/patents`, where one variant seems to have more third-party data

Content versus Metadata on the Internet

- (There are more examples in the paper, and more we didn't even include)
- The content/metadata distinction from 1979 doesn't translate well to the Internet
- Individuals do not see the behavior of Internet companies the way they (according to the Court) saw the phone company behave
- There are many parties that can behave differently under different circumstances, complicating any *a priori* analysis
- We need a new framework—but what?

Bird of the Day

Red-winged blackbird (non-breeding male), Central Park, March 23, 2021

CS@CU