Roadmap:
What Data, Exactly, is Out There?

Professor Doug Szajda

Before We Start, Some Preliminaries

- I am Professor Doug Szajda
  - Office: 212 Jepson Hall
  - Email: dszajda@richmond.edu
  - Office phone: 804-287-6671
  - Webpage: http://www.richmond.edu/~dszajda
  - My research area is computer networks and security
- Our Roadmap class Program Assistant (PA) is Nicolas Munsen
  - Nicolas' email: nicolas.munsen@richmond.edu
  - Nicolas' cell: (919) 816-6388

Before We Consider "What",…

- Some data is definitely stored somewhere, some is only *possibly* stored somewhere
  - E.g., content of phone conversations — some are stored, but only if some agency has a reason for doing so (in theory — NSA and others have been known to perform some random trolling)
  - E.g., identifies of those you've conversed with *definitely* stored — you can see this on your cell phone bill
- Also, for some data, it is easy for an adversary to know where it is stored. For other data, this is not the case
  - E.g., some medical records (need to know who my physician or specialist is/was — unless she/he wrote me a prescription)

So, Let's Start With the Public Stuff

- An experiment from last summer:
  - A local high school student
  - About two hours
  - Not allowed to pay for access to information
  - Not allowed to hack into any sites
  - the task: find out all you can about me
    - Note: because I'm just slightly older than you, there is a bit more info out there about me
    - BUT: when I was born, all of this information was NOT digitally recorded. And yet….

He Found the Following

- The date, time, and hospital of my birth, and who the physician was
- The names of my parents
- All of my addresses and phone numbers going back more than 30 years
  - That is well over a dozen addresses and numbers
- When I purchased my current house (the only house I have purchased) and what I paid for it, how many square feet it is, how many bedrooms and baths, the year it was built, when it was remodeled, the names (and some phone numbers) of all the former owners
  - But it did list my occupation as building and grounds cleaning

He Found the Following

- The names of all the schools I have attended, and the dates I attended them
  - Including the dates I received my degrees, if applicable
- All of the places I have taught (including high schools)
  - There are quite a few of them
- Of Course: All of my papers, grants, etc.
  - But these are easy
- The names of my siblings and children
  - And where my children attend school?

So What Else is Out There, but Perhaps not Public?

- Pretty much EVERYTHING about you!

So What Else is Out There, but Perhaps not Public?

- All birth related information
  - When, where, parents names, doctors
- Social security numbers
- All bank account/credit card/ license numbers
- Every purchase you/your family have ever made, unless it was paid in cash
  - And even that is there if you used a "frequent buyer card"
  - Consider the ramifications of this

So What Else is Out There, but Perhaps not Public?

- All text messages (who you texted and when)
  - Content of any individual text may or may not be saved somewhere (other than the party who received text)
    - NEVER text questionable pics (more on this later)
- All emails (who you emailed and when)
  - Again, content may or may not be saved
- All calls (when you made them and to who), whether land line or cell phone
  - Again, content may or may not be saved
  - Note that cell phone records allow one to track your location (ALL the time! Your cell phone off? Is your friend's on?)

So What Else is Out There, but Perhaps not Public?

- Skype calls
  - Content is encrypted, so private?

So What Else is Out There, but Perhaps not Public?

- Skype calls
  - Content is encrypted, so private?
  - Not so fast, my friend:

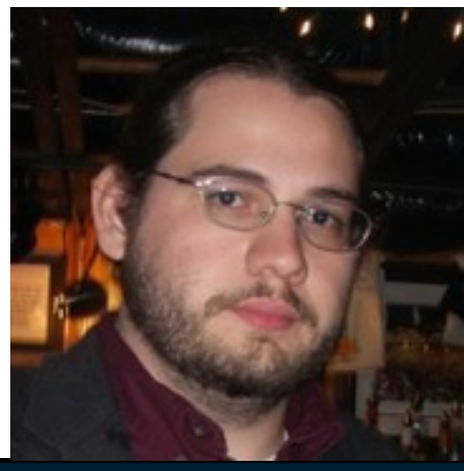
Phonotactic Reconstruction of Encrypted VoIP Conversations:
Hookt on fon-iks

Andrew M. White* Austin R. Matthews*† Kevin Z. Snow* Fabian Monrose*
*Department of Computer Science †Department of Linguistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
{amw, kzsnow, fabian}@cs.unc.edu, armatthe@email.unc.edu

***Abstract*—In this work, we unveil new privacy threats against Voice-over-IP (VoIP) communications. Although prior work has shown that the interaction of variable bit-rate codecs and length-preserving stream ciphers leaks information, we show that the threat is more serious than previously thought. In particular, we derive approximate transcripts of encrypted VoIP conversations by segmenting an observed packet stream into subsequences representing individual phonemes and classifying those subsequences by the phonemes they encode. Drawing on insights from the computational linguistics and speech recognition communities, we apply novel techniques for unmasking parts of the conversation. We believe our ability to do so underscores the importance of designing secure (yet efficient) ways to protect the confidentiality of VoIP conversations.**

ciphers for encryption—interact to leak substantial information about a given conversation. Specifically, researchers have shown that this interaction allows one to determine the language spoken in the conversation [55], the identity of the speakers [2, 41], or even the presence of *known* phrases within the call [56].

Rightfully so, critics have argued that the aforementioned threats do not represent a significant breach of privacy. For example, the language of the conversation might easily be determined using only the endpoints of the call—a call from Mexico to Spain will almost certainly be in Spanish. While the identification of target phrases is more damning, it still requires the attacker to know (in advance) what she



So What Else is Out There, but Perhaps not Public?

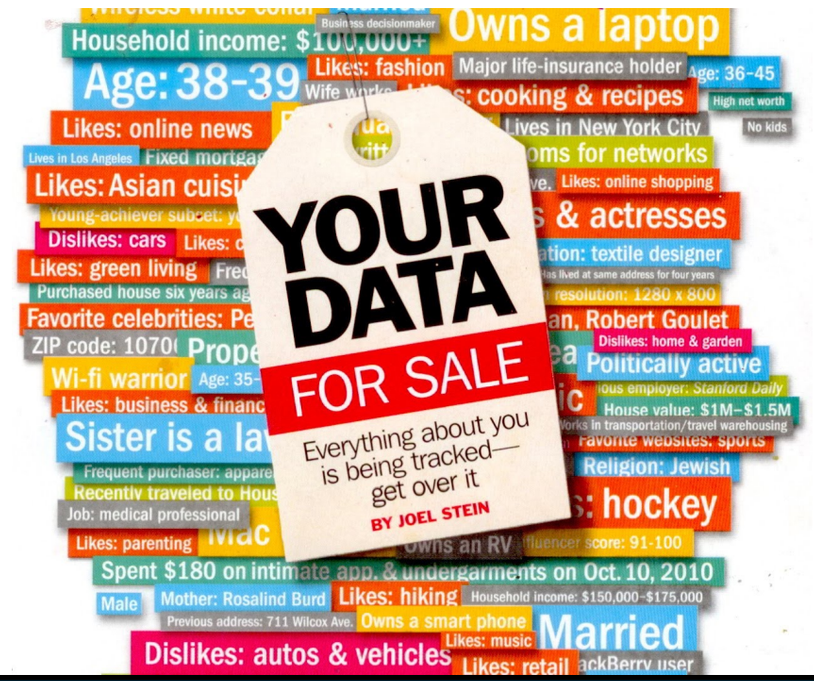
- What TV shows you watch in your home, when you watch them
- What movies you watch in your home, when you watch them (either via FiOS or the like, netflix, or redbox, etc.)
- What video game consoles you own, what videogames you own, who and when you play them with
- All books that you own
  - Certainly ebooks, but you did pay for the hardcopy books with credit/debit card, did you not? Or perhaps you paid cash but used your B & N member card?
  - Or maybe you give this info away for free on Goodreads?

So What Else is Out There, but Perhaps not Public?

- All medical records
  - Every prescription you have ever received and every surgery/ procedure you've undergone
    - Some might be embarrassing or damaging (e.g., sexually transmitted disease)
  - Now: records of all doctor visits
  - Records of all visits to counselors
  - Yes, this is all legally protected. But it exists. And not all who want access to it are people who obey the law
    - And who manages the protection of this stuff anyway?

So What Else is Out There, but Perhaps not Public?

- All legal records
  - Some are matter of public record (e.g., house sales) others are available upon request (depositions in divorce proceedings (ouch!))
  - Others are not, and are protected by attorney-client privilege
  - But again, they are stored digitally (and again, who protects them?)



So What Else is Out There, but Perhaps not Public?

- All web sites you have visited (and what you click on when you visit them)
  - Really doesn't matter what browser you use and what sites you visit — you are constantly being tracked
  - This information is shared and used to target ads…
    - So maybe you don't mind. But do you really want to receive targeted ads for itchy butt cream?
  - …or worse: Perhaps you are a member of some websites that might be damaging or embarrassing (e.g., Ashley Madison)
    - Better make sure no one with an agenda hacks them and releases the email addresses of everyone who has an account

So What Else is Out There, but Perhaps not Public?

- Voter registration records and donations to political parties
  - Oh, and have you signed an Internet petition lately?
- iCloud (and the like)
  - All of your photos
    - Which discloses info on significant others, friends, family, vacations, that great party where you are shown chugging (future employers love that)
  - All of your contacts
    - All on by default

But Why Make Others Work For this Info?

But Why Make Others Work For this Info?

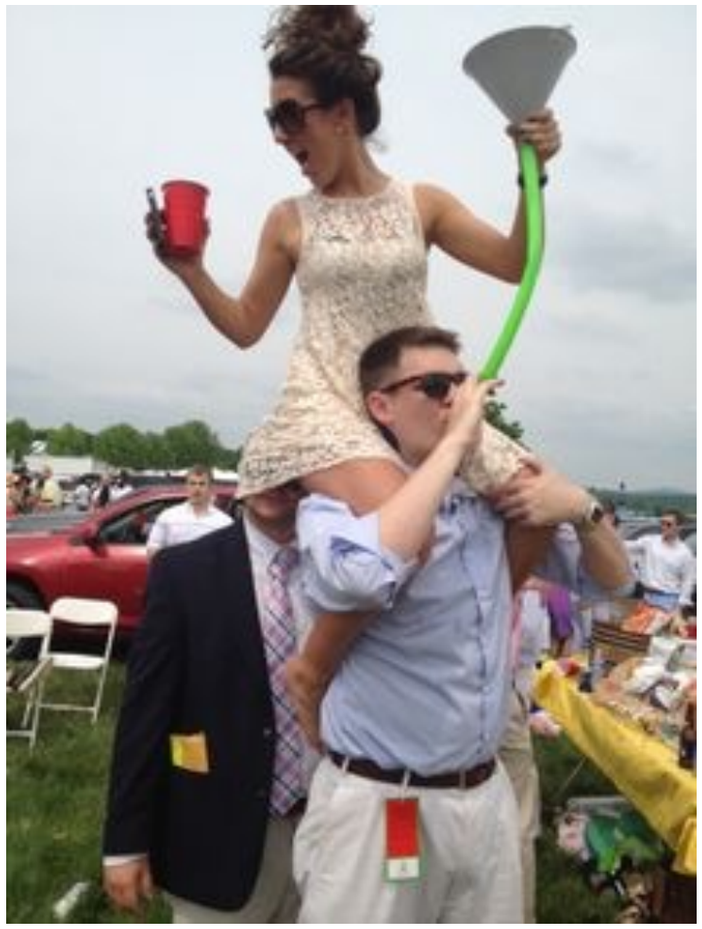
- I do not know Jim Smith
  - Searched on it because Jim Smith is a very popular name
  - Looked at this because it turns out Jim went to the same high school I did (and it turns out he even knows some of the same folks from that high school that I knew)
- I now know a list of lots of his good friends
- I know where he goes on vacation (because he uses Check-Ins or because I can look at his photos)
- I know a lot of his views on things, because I can see his friends' comments and his responses to them

But Why Make Others Work For this Info?

- I know for example, that he likes poetry and country music
- I could reconstruct his friend graph and probably find out who all of his closest friends are
- But I'm doing this for this class. Why would someone else do it?
  - Perhaps Jim has applied for a job (and the employer wants to make sure that Jim has "the right stuff" for the job — e.g., political or religious affiliation (illegal), temperament, dedication, etc.)
  - A picture is worth a thousand words…

But Why Make Others Work For this Info?

- These ladies will have no trouble being treated as a professional, should they be hired (not at all implying this is right, just stating fact)



But Why Make Others Work For this Info?

- Same for these gentlemen

But Why Make Others Work For this Info?

- Same for these gentlemen

And by the way, just because YOU didn't post them, doesn't mean your friends didn't!

Who is Collecting This Stuff?

- The myth years ago was that hackers, etc, were young kids (see the movie "War Games")
- The truth is that most groups doing hacking these days are very well financed and are able to attract much talent
- Organized crime (Target hack was result of a Ukranian group)
- Foreign governments: China has a well established military hacking division
- Our own government: The NSA
- Lots and lots of businesses!

The NSA

- Note: I am not making here a judgement on whether such surveillance is or is not warranted. Simply reporting on what is being collected
- Terminology: *Metadata* - information about communication - time, parties involved, locations of parties, etc.
- Though many in the security community knew about some of this, the Snowden leaks were very illuminating, especially to the public
- Among the programs confirmed or revealed were…

The NSA

- Phone calls: NSA collecting the phone records of millions of Verizon customers (via secret court orders)
  - And virtually every other phone company — companies wanted to make these orders public
- PRISM: NSA program to infiltrate the servers of Google, Facebook, Microsoft, Apple, others
  - On related note, Juniper networks discovered that outside code had been placed into the servers it was shipping. NSA?

The NSA

- XKeyScore: NSA tool for learning virtually anything a user does on the Internet
- Efforts to weaken (crack) encryption and undermine Internet security
- Programs designed to surreptitiously remove from standards committees parties that advocate for strong crypto and strong Internet security
  - This is another chapter in a decades long story
  - Argue that strong crypto allows terrorists and other adversaries to communicate without surveillance

The NSA

- NSA cracks Google and Yahoo datacenter links
- Dishfire program: NSA intercepts 194 million text messages EVERY DAY!
  - Or at least it did before the Snowden leaks
    - NSA documents described this as a "gold mine to exploit" for all kinds of personal data
- And it's not just you — if you become a person of interest, they can check all people within 3 degrees of separation — all the people who know all the people you know

The NSA

- Check out this very informative article in The Guardian: http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1

Why Worry About Privacy?

- If you're famous, it's pretty clear why
  - See, e.g.,Florida lawsuit Bollea v. Gawker
    - You likely think of it as Hulk Hogan v. Gawker
- iCloud nude photo scandal in 2014
  - 26 celebrities involved

Why Worry About Privacy?

- This is a question I get a lot, especially from students (not so much from older folks)
- "If I'm not doing anything wrong, what do I have to hide?"
  - Misses the point: Information can be misused or stated out of context
  - Example: Susan Jones, a fine kindergarten teacher, wants to show her class the White House web site. She types in www.whitehouse.com. What she gets is most definitely NOT the white house web site (it's an "adult" site — the White House site is www.whitehouse.gov). If the school is monitoring her Internet use, they can truthfully say that she showed her kindergarten class a pornographic site.

A Better Example: Barry Ardolf

- Computer tech from Minneapolis, age 46, sentenced in 2011 to 18 years in prison
- New neighbors reported him to police after he kissed their 4 year old son on the lips. He decided to get back at them
- Hacked into neighbors wifi, and transmitted "content" in attempt to destroy career and professional reputation
  - Created fake myspace page for husband, posted pics of underage girl having sex with two underage boys — page bragged that since he is lawyer, he could get away with it
  - Sent same porn to husband's co-workers

A Better Example: Barry Ardolf

- Sent, through husband's email account, flirtatious emails to women who worked in wife's office
- Send emails from husband's yahoo email account to Vice-president Biden and other politicians, claiming "this is a terrorist threat" and "I swear to God I'm going to kill you"
  - Needless to say, Secret Service showed up at husband's office
- Finally, forensics team working for husbands law firm found email session where threats originated, which allowed FBI to get warrant to search Ardolf's house and computers, where they found the evidence

A Better Example: Barry Ardolf

- BUT: try to imagine the nightmare this put the couple through
- And consider what could "truthfully" be said about origination of content!
- Read for yourself: https://www.wired.com/2011/07/hacking-neighbor-from-hell/

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Privacy is a limit on government and private sector company power
  - The more someone knows, the more power they have
  - Personal data guides decisions, affects reputations, can be used to influence decisions and shape behavior. Can be dangerous in the wrong hands
- Privacy is about respecting individuals
  - If a person has a reasonable desire to keep some info private, there needs to be compelling reason to disclose that info
    - Yes, sometimes conflicts with other important values. But should not be shrugged off as unimportant.

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Reputation management - privacy allows people to have some control over their reputation
  - Ability to protect from unfair harm
  - Protecting reputation depends not only on managing falsehoods, but on managing some truths
    - People judge badly, in haste, out of context, without knowing entire story, and, frankly, hypocritically. Knowing ALL the truth does not necessarily lead to a more accurate judgement about someone.
- Maintain appropriate social boundaries
  - People establish boundaries from others in society — privacy helps manage these
    - Think "none of your business" and "too much information"

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Trust — the basis for many (all?) relationships
  - Breaches of confidentiality are breaches of trust.
  - Relationships with doctors, lawyers, counselors, business partners
- When trust is breached, many are reluctant to trust again

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Control over one's life
  - Personal data essential to many decisions made about us - loans, licenses, jobs, whether we are investigated by government, searched at airport, allowed to fly, what messages and content we see over Internet
  - Without knowledge of how it is used, ability to correct and amend it, to have a say in how it is used, and to be able to air legitimate grievances, we are helpless.
  - Freedom based on individual autonomy and control - can't have it when decisions being made in secret without our knowledge or participation

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Freedom of thought and speech
  - Privacy is key to freedom of thought
    - Being watched can deter us from exploring ideas that are not considered mainstream (see examples from the former Soviet Union)
    - Person might want to explore ideas that are outside of societal or family norms, or ideas not popular among colleagues
- Freedom of social and political activities
  - Key component of political association is ability to do so with privacy
    - Ballots are kept secret to allow people to vote their conscience free from coercion

Why Worry About Privacy? (thanks to Daniel Solove, GW Law School)

- Ability to change and have second chances
  - People change and grow — the ability to move beyond past mistakes (within reason) is facilitated by privacy
  - Within reason, this should be encouraged, because positive growth is something that is good for society
    - You might appreciate this one more when you are older
- Not having to explain or justify yourself and your decisions
  - Many of the things we do, if judged by someone lacking complete information, may seem foolish, embarrassing, and/or reckless — do we want a society where we have to be prepared to justify all of these?

Why Worry About Privacy?

- In purely concrete terms, your personal information is worth quite a bit of money
  - There are many companies whose sole goal (and means of generating revenue) is learning information about you
  - Once they have that information, you no longer have any control over it (who sees it, how it is used, etc.)
  - So what you get for signing that information over to them (perhaps by agreeing to provide info for an Internet coupon or free game or some other treat) is nothing but the ability to have more targeted ads thrown your way

But the Internet will Crash and Burn!

- It won't. Yes, the ability to surf some sites for free is made possible by the revenue generated by targeted ads
  - But many would choose to have these ads in any case
  - Privacy advocates only seek to have web tracking default to otp-in (you must choose to allow it) as opposed to opt-out (where you must choose to disallow it)

- My opinion, for what it's worth
  - I like states like Vermont and Maine, where I'm not assaulted by huge billboards while driving down the highway
    - I see enough ads as it is — don't need more when I'm driving…
    - …or when surfing the internet (especially not when I pay to access a site)

But the Internet will Crash and Burn!

- It won't. Yes, the ability to surf some sites for free is made possible by the revenue generated by targeted ads
  - But many would choose to have these ads in any case
  - Privacy advocates only seek to have web tracking default to otp-in (you must choose to allow it) as opposed to opt-out (where you must choose to disallow it)

- My opinion, for what it's worth
  - I like states like Vermont and Maine, where I'm not assaulted by huge billboards while driving down the highway
    - I see enough ads as it is — don't need more when I'm driving…
    - …or when surfing the internet (especially not when I pay to access a site)

But Prof S, What does Digital have to do with it?

- After all, embarrassing pictures are not a new thing, nor are companies keeping purchasing records, etc.
- Well, yes and no. Yes: the records existed, usually in paper form.
  - Paper form is difficult to copy, and stealing it in a manner that is undetected requires copying it
  - And you have have physical access to the file in order to copy it.
  - And because physical access is required, sharing of this information between companies is not easily accomplished
    - There is lots of it. Sharing it requires moving it or copying it. This is expensive and a logistics nightmare. It's not practical.

But Prof S, What does Digital have to do with it?

- Digital data is easily and inexpensively copied, transferred, and stored
- A skilled adversary will likely be able to steal it from anywhere in the world!

- Sometimes even if the storage device is not attached to any network (see Stuxnet)
- And since copying it is trivial, it can be stolen without leaving a trace, and easily stored once stolen
- Image at left: server with 725 TB storage capacity
  - That's more than four times the memory required to store all the books every written

But Prof S, What does Digital have to do with it?

- But even if it's not stolen, digital information changes the game
- The ease of sharing and storage means that lots of organizations share data (and most of these have multiple backups of the data)

- This has the effect of concentrating the data into locations that are, to an adversary, very high reward!
  - So that "lock" icon on your browser is useless. Crooks aren't going to snoop on your purchase from B & N. They are going to attack the entire B & N database in the hopes of collecting many millions of credit/debit card numbers.
  - But there's one more important issue here…

Remanence: Digital Data Lives Forever!

- Because digital data is so easy to copy and share, it has a virtually unlimited life span. And you have no idea who has it or how many copies they have. So that embarrassing facebook pic of you — it will still be out there somewhere long after you are gone.
  - In fact, the task of securely deleting data is so difficult that is it a research area in it's own right
  - Some progress has been made on creating data that self-destructs, but that data must be created within specific experimental systems, and even then, there is absolutely no guarantee that all copies of some given data item have been destroyed
    - And in fact, no system of any kind can guarantee this

Remanence: Digital Data Lives Forever!

- Going to build a system that prevents files from being copied?
  - Better make sure the machines used to view the files don't allow screen shots
- Going to disallow screen shots?
  - Better make sure the people viewing the files don't have cell phones that take pics?
- Going to prevent people from having cell phones and other devices while viewing files?
  - Better make sure the person viewing the file doesn't have a photographic memory (these people do exist)

Conclusion

- There is a lot of digitally stored personal data about you
- Protecting it is important
  - Unless you just want to give it all away
- Digital storage is not at all the same as paper storage
  - It raises a lot more tough questions and concerns regarding privacy
  - And it never goes away
- Next: What methods exist to protect data
- Then: How do adversaries attack