Computer Security (COM-301)

Privacy

Slides created by Carmela Troncoso
Some slides/ideas adapted from: George Danezis, Bart Preneel, Claudia Diaz, Seda Guerses

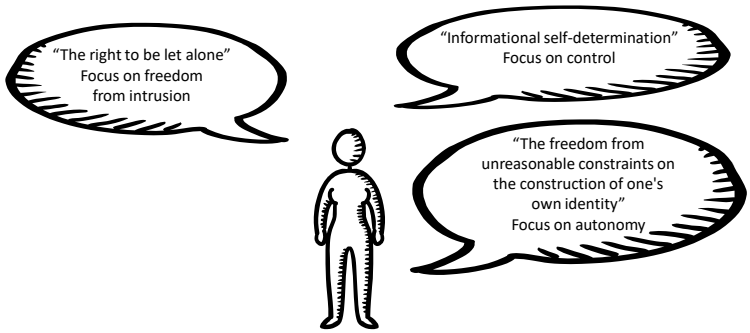
Goal of this lecture

Understanding:

- Privacy is not solely a local, individual-oriented problem. It is a *global property*
- There are different conceptions of privacy depending on the adversary model
- Depending on the adversary model one relies of different Privacy Enhancing Technologies: different protection degree
- Privacy requires to protect information beyond content: The need to protect meta-data

What is privacy

Abstract and subjective concept, hard to define
Dependent on cultural issues, study discipline, stakeholder, context



Privacy is a very abstract concept. In general is very subjective and depends on our culture and education; and most of the times also on the context: A sentence "I went for dinner with the girl I like" may not be private among friends, but may feel very private in a work environment.

There exist definitions in the literature mostly from legal and sociological perspective. Three very much used definitions are the following?

- "The right of be let alone" (Warren and Brandeis, 1890) focuses on freedom of intrusion. This is a quite US-centric definition at the time motivated by the beginning of photography and other technologies that could threaten intimacy. The idea is that people have the right to keep their own information without anyone having the right to look into it or publish it in any form.

- "Informational Self-determination" (Westin, 1967: "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others"). This definition focuses not on the idea that no-one should get access to information, but users should have the right to decide how to share their information.

- Freedom to construct one’s identity (Agre and Rotenberg 2001). This privacy conception is similar to Westin, but it goes beyond control to say that it is not only about what and when information is revealed, but about how do we expect people to process this information and form an idea of ourselves. In other words privacy helps us project the persona we want to different people (e.g., we are different at home with our family, than at the bar with friends, than at work with colleagues).

However, none of these can be directly used to design systems. They are hard to reason about when it comes to how to implement technology that helps supporting them.

In the legal system one can see Privacy as a right, and Confidentiality as a duty.

In a technical system, you cannot program this kind of “privacy." You have to formalize specific measurable attributes (properties) that result in privacy.
So when a security architect says "Privacy," they usually mean specific system states like Unlinkability, Anonymity, etc...

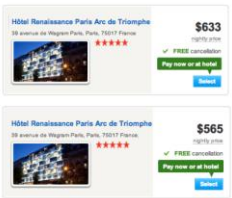
The context: Availability of data

Intelligent data-based applications

- Recommendation systems
  - Movies (Netflix)
  - Products (Amazon)
  - Friends (Social networks)
  - Music (Spotify, iTunes)
- Location based services
  - Friend finders
  - Maps
  - Points of interest
- Health monitoring
- Children/Elderly trackers
- Smart metering
- Intelligent buildings

Individual applications are legitimate

Together they become a cheap
SURVEILLANCE INFRASTRUCTURE



In our life we use daily data-driven applications for entertainment, shopping, or as a means to manage our relationships.

Each of these applications, individually, get a legitimate amount of data. When all of these data are put together, these applications become a very cheap surveillance infrastructure that can be used to learn every aspect of our life.

Once information is out there, it can be utilized, for instance to discriminate

100K users installed CA Facebook App

enabled **COLLECTING PERSONAL DATA** of 87+ million

public profile, page likes, birthday and current city

creation of **PROFILES** of the subjects of the data

TARGETED ADVERTISEMENTS during the US elections

A small number of users installed an application that “exploited” Facebook APIs to not only gather information about them but also their contacts, 87 million of people – two orders of magnitude higher.

This information, seemingly innocent (even somehow public) enabled the creation of precise profiles about these people and their preferences. This profiling helped the political campaign in the 2016 united states elections helping to decide who were the people that could be swayed to change their vote and what were the best advertisements for that purpose.
Political profiling is not a new thing, many years ago political parties were actively networking and using their local anchors to hierarchically produce understandings of different communities and how they vote: if you were an Irish Catholic in Boston or a Union worker in Manchester your vote was assumed to be highly predictable.
By just polling a few of these big classes, political figures would run their campaigns.

Cambridge Analytica pushed that quantitatively to the next level. Data analytics for political campaigning + try to influence/trigger behavior

https://www.reuters.com/legal/facebook-parent-meta-pay-725-mln-settle-lawsuit-

relating-cambridge-analytica-2022-12-23/

The context: Availability of data

Intelligent data-based applications

- Recommendation systems
  - Movies (Netflix)
  - Products (Amazon)
  - Friends (Social networks)
  - Music (Spotify, iTunes)
- Location based services
  - Friend finders
  - Maps
  - Points of interest
- Health monitoring
- Children/Elderly trackers
- Smart metering
- Intelligent buildings

Individual applications are legitimate

Together they become a cheap
SURVEILLANCE INFRASTRUCTURE

We need privacy!

But what about security!!?!?!?!

It is clear that to avoid surveillance, we need to improve online privacy, but what about security? Is it the case that privacy is at odds with security?

Common belief: we need to tradeoff security for privacy!

"For National Security surveillance is good and privacy is bad"

(Surveillance == Security) == True ??

*Surveillance may be not **effective***: smart adversaries evade surveillance

criminals use Telegram, Threema, Signal,... but average users do not!!

*Surveillance tools can be **abused***: lack of transparency and safeguards

NSA spying on Americans, Spanish ministry spying independentist politicians, Companies

*Surveillance tools can be **subverted** for crime*

Greek Vodafone scandal (2004-2005): "someone" used the legal interception functionalities (backdoors) to monitor 106 key people

7

From Wikipedia Greek Watergate: "The Ericsson switches used by Vodafone Greece were compromised and unauthorized software was installed that made use of legitimate tapping modules, known as "lawful interception", while bypassing the normal monitoring and logging that would take place when a legal tap is set up"

Privacy IS a security property

For individuals
protection against profiling and manipulation.
protection against crime / identity theft

For companies
protection of trade secrets, business strategy, internal operations, access to patents

For governments / military
protection of national secrets, confidentiality of law enforcement investigations, diplomatic activities, political negotiations

Privacy is actually a security property.

For **individual**:
As a lot of online security mechanisms (e.g., security questions to recover passwords) are based on private information, privacy is needed to protect online accounts -> i.e., privacy becomes a security mechanism.
Privacy is also important to control who gets access to our information. This is key to avoid profiling which in turn protect ourselves from manipulation (e.g., via personalized advertising). [See Facebook/Cambridge Analytica scandal as a paradigmatic example of the dangers stemming from an entity learning too much about social network users]

Example: saying online "I am going on vacation" -> people could decide to steal you

The fact that you say "great evening with my friend John."-> John is not at home. Someone could steal them.

For **companies:**
Digital interactions may reveal a lot about business decisions, e.g., mergers between

companies, launching of a new product.

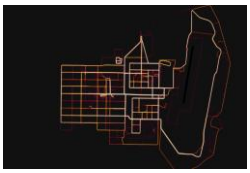
For **Governments**:
As for companies, digital traces reveal a lot about intentions such us who is being investigated by the police, which countries are talking which each other, which parties within a country are negotiating, etc.

Privacy IS a security property

INFRASTRUCTURE IS SHA

Individuals, Industry, and Governments use the same applications

Denying privacy to some is denying privacy to all!!



Directly
(Cloud-based services, Industry 4.0, Blockchain)

Indirectly
(employers are users)

One of the main reasons why privacy is as important for companies and governments are for individuals is because:

1) We all share the same infrastructure. There is only one Internet where all communication happens, and there are so many cloud systems where both companies and users do their computations or host their digital content.

2) Also, employees are users, and through their use of digital services may reveal information about their jobs or employers. The image on the right is a small piece of the heatmap of tracks followed by runners published by the social network Strava. This heatmap revealed the location of secret US military basis through the paths that soldiers use to run (https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases)

and Privacy is important for society

Daniel Solove, Prof. of Law

"Part of what makes a society a good place in which to live is the extent to which it allows people freedom from the intrusiveness of others. **A society without privacy protection would be suffocation**"

Not so much Orwell's "Big Brother" as Kafka's "The Trial":

"...a bureaucracy with inscrutable purposes that uses people's information to make important decisions about them, yet denies the people the ability to participate in how their information is used"

"The problems captured by the Kafka metaphor are of a different sort than the problems caused by surveillance. They often do not result in inhibition or chilling. Instead, they are problems of information processing—the storage, use, or analysis of data—rather than information collection."

"...not only frustrate the individual by creating a sense of helplessness and powerlessness, but they also affect social structure by altering the kind of relationships people have with the institutions that make important decisions about their lives."

may become

ONE RING TO RULE THEM ALL

And surveillance is not only a problem because of the learning of secrets, but because when people feel observed they change their behavior. It so happens that if when citizens feel that their behavior may have unforeseen consequences they start changing their behavior and relationships with power structures to try to compensate and influence their decisions.

Bring back the idea of panopticon.

What is privacy in Privacy Enhancing Technologies

PETs

3 different types of PETs depending on ...

the concerns they address

their goals

their challenges and limitations

Gürses, Seda, and Claudia Diaz. "Two tales of privacy in online social networks." IEEE Security & Privacy 11.3 (2013): 29-37.
Diaz, Claudia, and Seda Gürses. "Understanding the landscape of privacy technologies." Information Security Summit (2012): 58-63.
Danezis, George, and Seda Gürses. "A critical review of 10 years of privacy technology." Surveillance cultures: a global surveillance society (2010): 1-16.

In this course we will see that privacy takes a different in the different type of **Privacy Enhancing Technologies (PETs)**.

These are variations of the notion of privacy that can help when designing systems, and they differ from each other in:

- The *concerns* they address and *who* defines these concerns

- The adversarial model they aim at defeating (their *goal*)

- How far can they go in protecting privacy (their *limitations*)

We follow the nomenclature by Gürses and Diaz:
https://www.esat.kuleuven.be/cosic/publications/article-2270.pdf

1 – The adversary is in your **social** circle

CONCERNS - The privacy problem is defined by **Users**

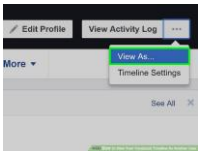
Technology brings problems

- "My parents discovered I'm gay"
- "My boss knows I am looking for other job"
- "My friends saw my naked pictures"

GOALS - Do not surprise the user

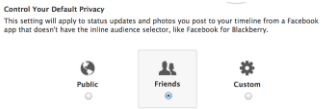
Two main approaches

- Support decision making
- Help identifying actions impact



Contextual feedback

Privacy nudges



Easy defaults

The first type of PETs we see in the class are termed "Social privacy" as they attempt to solve a "social" problem in which *users* worry about their social peers (friends, colleagues, acquaintances) learning about their private information through technology. The main fear is that this information can influence their relationships and opportunities.

The goal of social PETs is to help users with their use of technologies so that they are not surprised about how much others learn about them. There are two typical goals in these technologies:

- *Support decision making*: technologies that help users choose who can see what (e.g., default privacy policies).
- *Help identifying actions impact*: technologies that help users understanding how the information they put online may be seen and/or perceived by others (e.g., contextual feedback – which allows one to see how others see the information on the service as privacy mirrors; or privacy nudges – which use machine learning to predict how other users may perceive published texts or photos).

1 – The adversary is in your **social** circle

CONCERNS - The privacy problem is defined by **Users**

Technology brings problems

- "My parents discovered I'm gay"
- "My boss knows I am looking for other job"
- "My friends saw my naked pictures"

GOALS - Do not surprise the user

Two main approaches

- Support decision making
- Help identifying actions impact

LIMITATIONS

Only protects from other users: **trusted service provider**!
Limited by users' capability to understand policies
Based on user expectations – What if the expectations are null?

PETs for social privacy aim to protect users from other users, but do not consider the service provider as part of their adversarial model. The service provider is trusted with the data.

Also, these technologies are typically limited by the fact that users cannot really understand fine grained policies (i.e., users cannot deal with one policy per friend), thus policies are quite coarse and in general this ends up in being more data being shared.

Finally, recall that the goal of these technologies is to not surprise the user. If the user has no privacy expectation, these technologies do not need to do anything!

PETs for social privacy are very often implemented by industry, as they make users comfortable with the use of services, but still permit data collection.

1 – The adversary is in your **social** circle

CONCERNS - The privacy problem is defined by **Users**

Technology brings problems

- "My parents discovered I'm gay"
- "My boss knows I am looking for other job"
- "My friends saw my naked pictures"

GOALS - Do not surprise the user

Two main approaches

- Support decision making
- Help identifying actions impact

LIMITATIONS

Only protects from other users: **trusted service provider**!

Limited by users' capability to understand policies

Based on user expectations – What if the expectations are null?

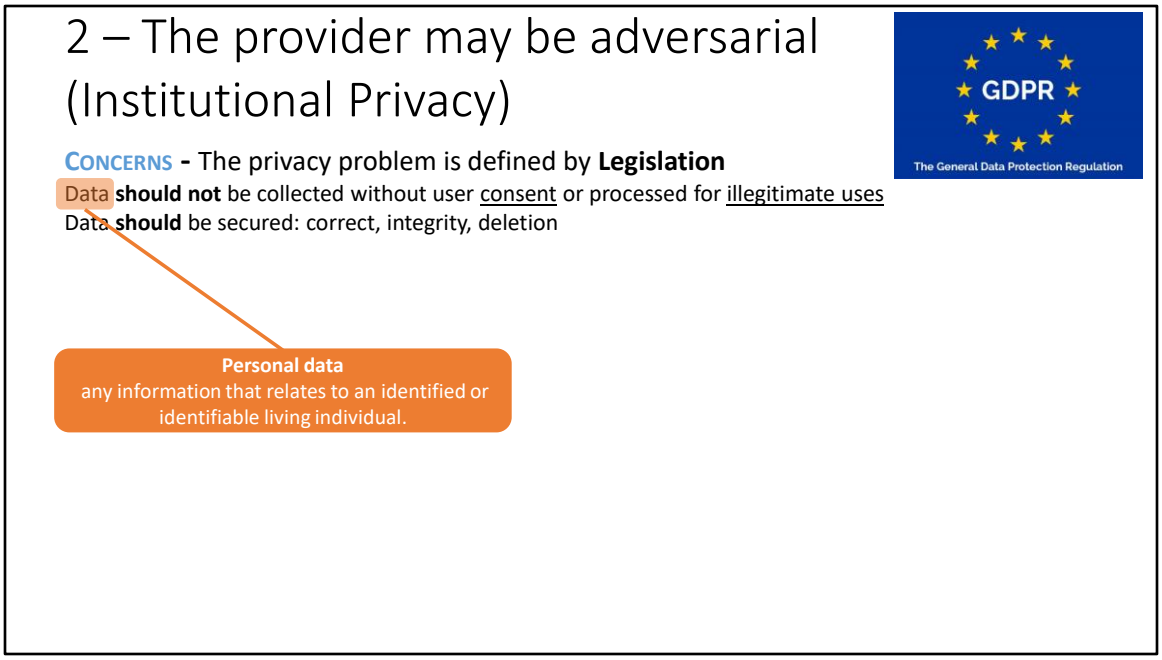
Common Industry approach
Make users comfortable

PETs for social privacy aim to protect users from other users, but do not consider the service provider as part of their adversarial model. The service provider is trusted with the data.

Also, these technologies are typically limited by the fact that users cannot really understand fine grained policies (i.e., users cannot deal with one policy per friend), thus policies are quite coarse and in general this ends up in being more data being shared.

Finally, recall that the goal of these technologies is to not surprise the user. If the user has no privacy expectation, these technologies do not need to do anything!

PETs for social privacy are very often implemented by industry, as they make users comfortable with the use of services, but still permit data collection.



The second type of PETs we see in the class are termed "Institutional privacy" as they are built to solve the privacy problem as they are defined by institutions through legislation and regulations (such as the General Data Protection Regulation – GDPR). Legislation only applies to **personal data**. These are data that relates to an identified individual (i.e., it is associated to this individual's name) or to an identifiable individual (i.e., it is not associated directly with the name but the data itself is enough to infer this name – e.g., learning the name of someone given their address and other points of interest).

The legislation focuses on data being only collected for legitimate purposes (with legitimate interest for a business, general interest, etc), or at least under *informed consent* (i.e., users must be informed about what information is going to be collected and how it is going to be processed and shared with third parties). Once data is collected, the legislation also mandates that they are secured, i.e., their integrity and correctness must be guaranteed; and that if deleted the deletion is secure (i.e., the data cannot be recovered under no circumstances).

They cannot collect data just because it might be useful later. They must have a specific legal reason for every piece of data they hold (e.g., "We need this address to

deliver the package" or "The user explicitly consented to this newsletter“, “the user want food for recommendation, so we collect all the restaurants they go to, to provide a better service”).
They must build systems that allow users to exercise their rights. If a user asks "What data do you have on me?" or "Delete me," you must be able to respond within 30 days.

Under the GDPR's "Accountability Principle“ a service provider is not "innocent until proven guilty." If a regulator knocks on their door, they are considered non-compliant unless they can produce the documentation to prove otherwise

2 – The provider may be adversarial (Institutional Privacy)

CONCERNS - The privacy problem is defined by **Legislation**

Data **should not** be collected without user consent or processed for illegitimate uses

Data **should** be secured: correct, integrity, deletion

GOALS – Compliance with data protection principles

informed consent
purpose limitation
data minimization
subject access rights

Preserving the security of data
Auditability and accountability

The goal of institutional PETs is to support data protection principles:

- *informed consent*: users must be informed about what information is going to be collected and how it is going to be processed and shared with third parties.
- *purpose limitation:* the data that is collected can only be processed for purposes in accordance with the goal of the application
- *data minimization*: the application must collect the minimal amount of data necessary for the provision of the service
- *Subject access rights:* users have the right to demand service providers to tell them what data has been collected about them, how it has been processed, and demand its correction or deletion.

  [users whose data is collected are called subjects in the regulation]
- *Data security*
- *Auditability and accountability*: companies must make sure that their collection and processing can be audited (i.e., they can prove what data they collected and what happened to the data); so that if something goes wrong they can be held accountable for the problem.

Technologies to help with these principles are:

- *Logging*: storing what actions have been performed on the data and by whom.

- *Access control:* naturally, it helps preventing unauthorized parties from accessing data, and it also helps identifying the principal that accesses the data to the logging.

Anonymization: is a technology that aims at decoupling data from the identity so that it is not considered personal data anymore. Once it is not personal data, it is **not** subject to the data protection regulation.

2 – The provider may be adversarial (Institutional Privacy)

CONCERNS - The privacy problem is defined by **Legislation**

Data **should not** be collected without user consent or processed for illegitimate uses

Data **should** be secured: correct, integrity, deletion

GOALS – Compliance with data protection principles

informed consent
purpose limitation
data minimization
subject access rights

Preserving the security of data
Auditability and accountability

Access control
Logging
Anonymization??

Wouldn't it be nice if... you could take a dataset full of personal data, and transform it into one with no personal data – while keeping all the value of the data?

Magic does not exist!
this **cannot** happen in general!

The goal of institutional PETs is to support data protection principles:

- *informed consent*: users must be informed about what information is going to be collected and how it is going to be processed and shared with third parties.
- *purpose limitation:* the data that is collected can only be processed for purposes in accordance with the goal of the application
- *data minimization*: the application must collect the minimal amount of data necessary for the provision of the service
- *Subject access rights:* users have the right to demand service providers to tell them what data has been collected about them, how it has been processed, and demand its correction or deletion.

  [users whose data is collected are called subjects in the regulation]
- *Data security*
- *Auditability and accountability*: companies must make sure that their collection and processing can be audited (i.e., they can prove what data they collected and what happened to the data); so that if something goes wrong they can be held accountable for the problem.

Technologies to help with these principles are:

- *Logging*: storing what actions have been performed on the data and by whom.

- *Access control:* naturally, it helps preventing unauthorized parties from accessing data, and it also helps identifying the principal that accesses the data to the logging.

Anonymization: is a technology that aims at decoupling data from the identity so that it is not considered personal data anymore. Once it is not personal data, it is **not** subject to the data protection regulation.

2 – The provider may be adversarial (Institutional Privacy)

CONCERNS - The privacy problem is defined by **Legislation**

Data **should not** be collected without user consent or processed for illegitimate uses
Data should be secured: correct, integrity, deletion

GOALS – Compliance with data protection principles

- informed consent
- purpose limitation
- data minimization
- subject access rights
- Preserving the security of data
- **Auditability and accountability**

LIMITATIONS

Never questions collection – assumes it is necessary
Trusted service provider! No technical measures to protect data from them
Limits misuse, but not collection (seven legal basis)
Limited scope (personal data != all data)

The PETs that support institutional privacy aim at protecting the data or its uses, but very rarely at diminishing collection. The only principle in this direction, data minimization, is typically enforced at the policy level and not by technology.

As social privacy PETs, these technologies do not really protect the user from the trusted provider, other than a posteriori punishment if the provider misbehaves.

It only applies to personal data, which can be used to profile individuals, but not to other data (e.g., city sensors) that are still useful to profile populations and make decisions to influence those populations as a whole.

PETs for institutional privacy are implemented by industry, as compliance is mandatory under many regulations (in particular in Europe, and for services that include European citizens regardless of where the services are).

Examples of PET:

Tools (like OneTrust or Cookiebot) that pop up when you visit a site are PETs designed

specifically for the "Informed Consent" goal. They record exactly what users agreed to and ensure data flows only to approved vendors.

Secure Multi-Party Computation:This allows multiple institutions (like hospitals or banks) to compute a result (e.g., "average salary") without any institution ever seeing the other's raw data.
This satisfies Data Minimization and Purpose Limitation because the provider never holds data they don't strictly need for the final calculation.

Federated Learning: Instead of a provider collecting all user data on a central server (which creates a privacy risk), an AI model is sent *to* the user's device.
The model learns locally and only sends back mathematical updates, not raw data.
This is a PET for Data Minimization, though it is hard to reason exactly on what information is contained in the gradient/mathematical updates.

3 – “Everyone” is the adversary (Anti-surveillance Privacy)

CONCERNS - The privacy problem is defined by **Security Experts**
Data is disclosed **by default** through the ICT infrastructure: **the adversary is anybody**
Concerned about: censorship, surveillance, freedom of speech,…

The third type of PETs we see in the class are termed “Anti-surveillance privacy” as they are built to solve the privacy problem that stems from the fact that the ensemble of applications and the infrastructure itself leak so much information that they become a surveillance infrastructure. These technologies *consider the infrastructure itself and the providers as the adversaries*.

The goal of these technologies is the disclosure of any information to anyone. This comprises both information explicitly revealed (such as content) and revealed implicitly through the use of infrastructure (such as IPs, location, etc). A main goal of minimizing information disclosure is to reduce the amount of trust to preserve privacy on other entities. This in many occasions includes not only reduce the amount of information on one entity but also sharing information among entities in such a way that no individual entity can breach privacy. *(this is yet another instance of the separation of privilege principle in use).*

3 – Everyone is the adversary (Anti-surveillance Privacy)

CONCERNS - The privacy problem is defined by **Security Experts**

Data is disclosed **by default** through the ICT infrastructure: **the adversary is anybody**

Concerned about: censorship, surveillance, freedom of speech,…

GOALS – Minimize

Default disclosure of personal information to anyone - both explicit and implicit!

Minimize the need to trust others

The third type of PETs we see in the class are termed "Anti-surveillance privacy" as they are built to solve the privacy problem that stems from the fact that the ensemble of applications and the infrastructure itself leak so much information that they become a surveillance infrastructure. These technologies *consider the infrastructure itself and the providers as the adversaries*.

The goal of these technologies is the disclosure of any information to anyone. This comprises both information explicitly revealed (such as content) and revealed implicitly through the use of infrastructure (such as IPs, location, etc). A main goal of minimizing information disclosure is to reduce the amount of trust to preserve privacy on other entities. This in many occasions includes not only reduce the amount of information on one entity but also sharing information among entities in such a way that no individual entity can breach privacy. *(this is yet another instance of the separation of privilege principle in use).*

3 – Everyone is the adversary (Anti-surveillance Privacy)

CONCERNS - The privacy problem is defined by **Security Experts**
Data is disclosed **by default** through the ICT infrastructure: **the adversary is anybody**
Concerned about: censorship, surveillance, freedom of speech,…

GOALS – Minimize
- Default disclosure of personal information to anyone - both explicit and implicit!
- Minimize the need to trust others

LIMITATIONS
Privacy-preserving designs are narrow – very difficult to create "general purpose privacy"
Usability problems both for developers and users
- how the @$%&#$Ŷ& do I program this?
- performance hit
- unintuitive technologies

Industry lacks incentives

Anti-surveillance PETs can enable services with minimal information disclosure (e.g., processing data in the encrypted domain), but these technologies can typically make one thing at time. It is hard to create general purpose technologies that can be applied to any problem.

They are also hard to use. For developers because they rely on very complex mathematics, and result on high overhead in bandwidth and/or computation time; and for users because they are many times non-intuitive as they enable operations that cannot be done in the physical world (e.g., anonymous authentication in which one can prove one attribute – the equivalent would be to show your passport but only reveal that the date of birth is before 18 years ago).

Finally, industry has no incentives to roll out these technologies that limit the amount of data that they can collect about users.

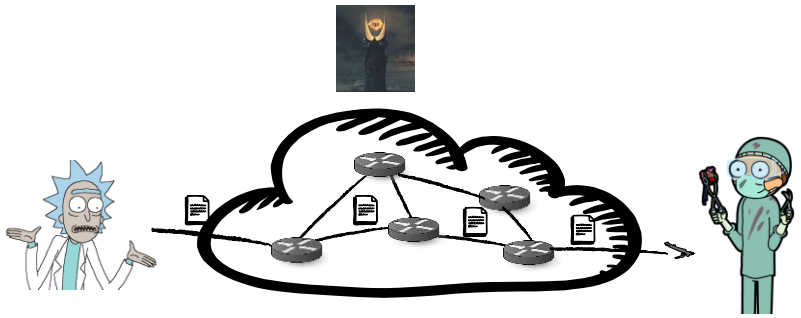
Some Privacy Technologies

Everyone is the adversary

Once a message is on the network, it can be accessed by anyone that has access to the medium.

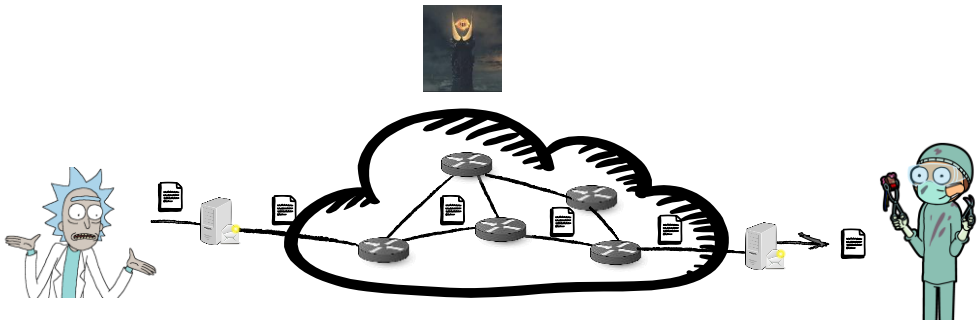
Also, once data is stored at the provider it can be accessed by anyone that can request it: the provider itself, but also law enforcement agencies with a subpoena, or intelligence agencies that have deals with service providers (https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data)

End to End Encryption



End to End Encryption

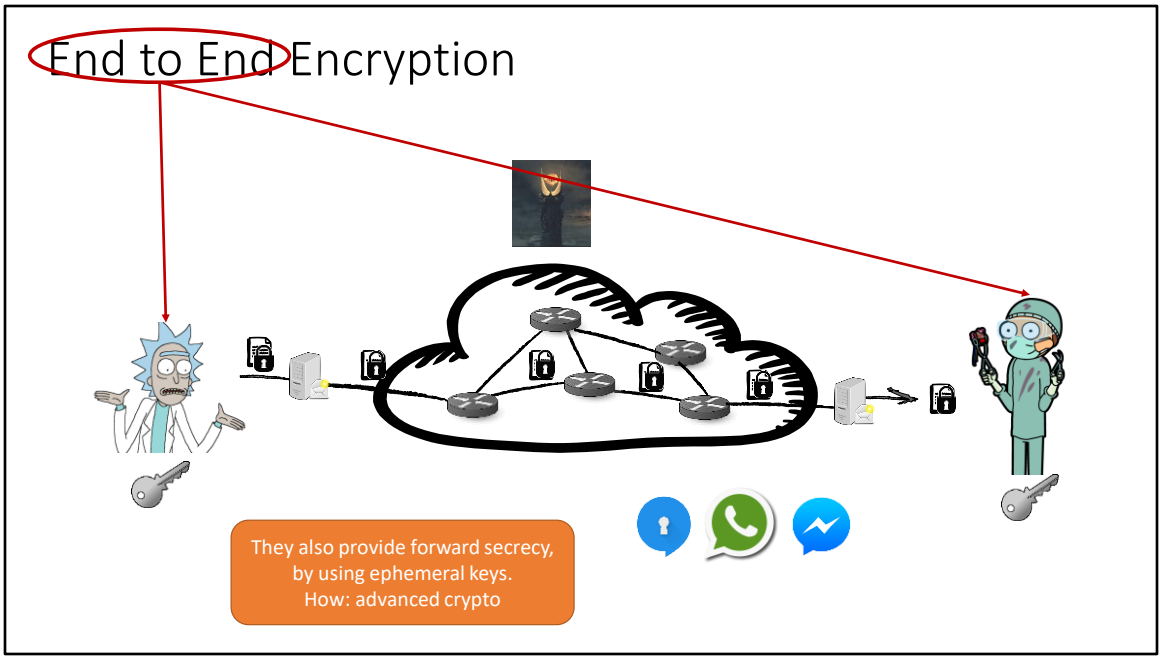
What is an End?



Cryptography → Confidentiality!
(and integrity and authenticity)

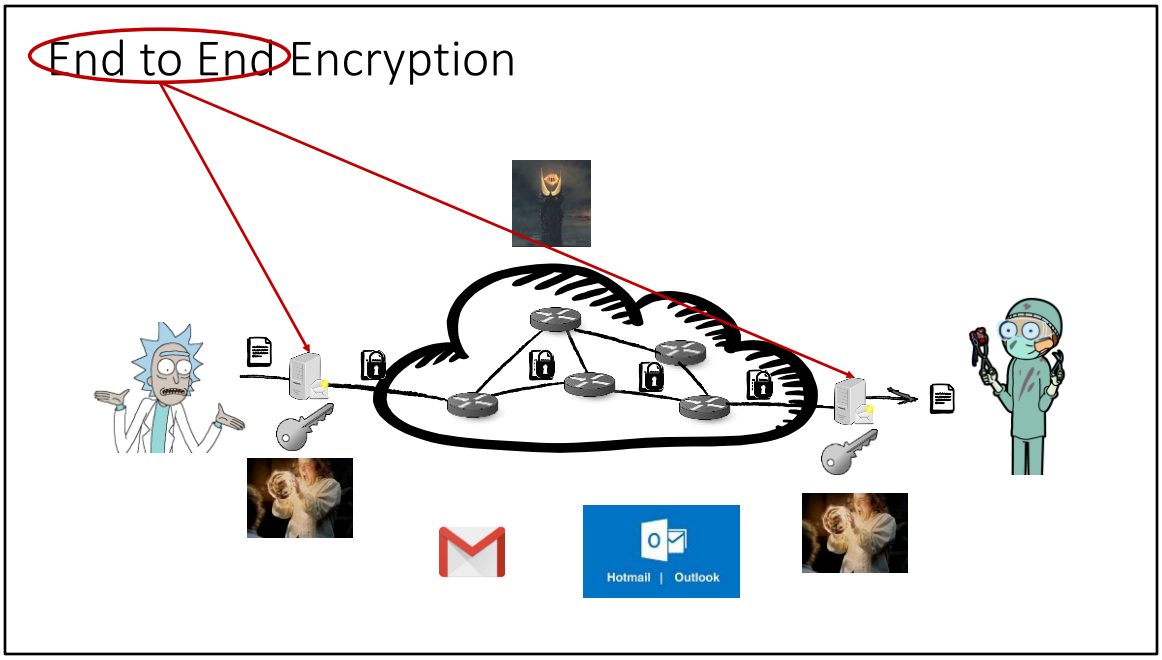
A solution is using cryptography to achieve confidentiality of the content.

We see a lot the term **end to end encryption**. The protection provided by end to end encryption depends very much on what is considered an end.



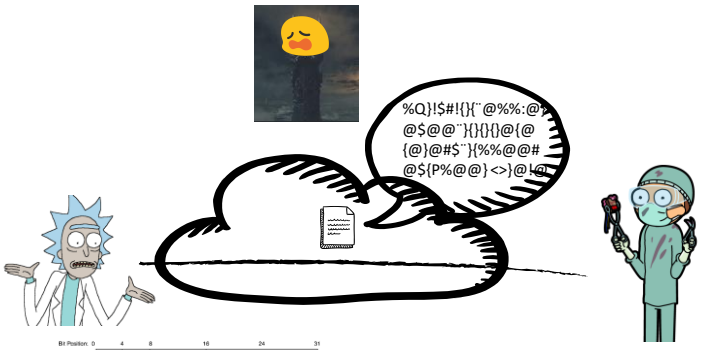
When the ends are the user devices (i.e., encryption is done on the devices that *store the encryption keys*) it provides protection from any entity on the path including the service provider.

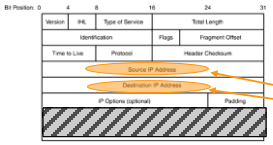
This is implemented, for instance, by current messengers such as Whatsapp, Signal or Telegram. These applications also use (or have the options to use) ephemeral keys for the messages such that even if the phone is compromised at some point, previous messages cannot be decrypted.



When the ends are the service provider servers, and end to end is implemented as using TLS connections between clients and server, and servers, the protection is only against third parties that can observe the network. However, the data is in the clear and can be accessed by the provider as well as anyone that can force the provider to disclose the data.

But we can encrypt! What is the problem?



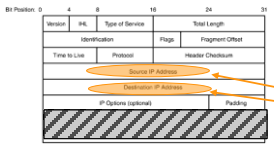


IPv4 Header (RFC 791, 1981)

Same for Ethernet, TCP, SMTP, IRC, HTTP, ...

Even when the content is encrypted much of the other information, such as source and destination of the messages, are available to the observer. The slide shows the example of IP, but this is common to all protocols underlying digital communications.

The problem is Traffic Analysis

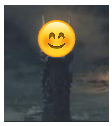


IPv4 Header (RFC 791, 1981)

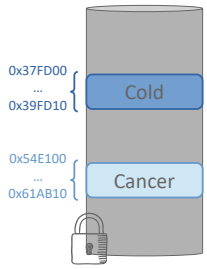
Same for Ethernet, TCP, SMTP, IRC, HTTP, ...

The availability of communication information (also known as meta-data) enables the adversary to perform **traffic analysis.**

Other metadata is also sensitive!!



Implicit data is as important as explicit data!



The address where data is stored may reveal information about the content.
Example: medical database with patients with mild and severe diseases in different locations

The address where an action happens may reveal information about the action / user.
Example: sending a message from an Oncologist clinic reveals information about the sender

Analysis of metadata to circumvent encryption's protection can be applied to other data than traffic data: e.g., which addresses are accessed inside a system may provide information about the content of the memory, or the location of a user when using a digital service can reveal a lot about the nature and content of the communication.

Think of an insurance providers, if they could get that metadata.

Traffic analysis

Wikipedia: traffic analysis is the process of intercepting and examining messages in order to deduce information from patterns in communication

Making use of "just" traffic data of a communication (aka metadata) to extract information (as opposed to analyzing content or perform cryptanalysis)

MILITARY ROOTS

M. Herman: "These non-textual techniques can establish **targets' locations**, order-of-battle and **movement**. Even when messages are not being deciphered, traffic analysis of the target's Command, Control, Communications and intelligence system and its patterns of behavior provides indications of his **intentions** and **states of mind**"

WWI: British troops finding German boats.

WWII: assessing size of German Air Force, fingerprinting of transmitters or operators (localization of troops).

NOWADAYS

Diffie&Landau: "Traffic analysis, not cryptanalysis, is the backbone of communications intelligence"

Stewart Baker (NSA): "metadata **absolutely tells you everything about somebody's life**. If you have enough metadata, you don't really need content."

Tempora, MUSCULAR → XkeyScore

Herman, Michael. Intelligence power in peace and war. Cambridge University Press, 1996.
Diffie, Whitfield, and Susan Landau. Privacy on the line: The politics of wiretapping and encryption. MIT press, 2010.
http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded

Traffic analysis is the process of analyzing metadata associated to communications such as the identity of the participants; when, how often, and for how long they talk; where they are; which device they use (e.g., office computer vs. mobile phone).

Traffic analysis has been used for long in military contexts. Nowadays, it is used for law enforcement, and also by systems to learn more about users (more about this in the CS-523 course).

Richness of Metadata: Browser fingerprinting

- amiunique.org analyzes your browser's configuration (e.g., screen resolution, fonts, timezone, user agent) and compares it against a massive database of collected fingerprints to calculate how distinct your device is from everyone else's
- It reveals which specific metadata points make your browser identifiable
- Proves that you can be tracked across the web even without using cookies or logging in

We need to protect the communication layer!
Why **anonymous communications?**

If you are a cyber-criminal!
- DRM infringement, hacker, spammer, terrorist, etc.

But, also if you are:
- Journalist
- Whistleblower
- Human rights activist
- Business executive
- Military/intelligence personnel
- Abuse victims

Or you want to…
- Avoid tracking by advertising companies
- Protect sensitive personal information from businesses, like insurance companies, banks, etc.
- Express unpopular or controversial opinions
- Have a dual life
  - A professor who is a pro in LoL!
- Try uncommon things
- …

https://www.eff.org/deeplinks/2013/10/online-anonymity-not-only-trolls-and-political-dissidents
http://geekfeminism.wikia.com/wiki/Who_is_harmed_by_a_%22Real_Names%22_policy%3F

A family of technologies that protect traffic metadata to avoid traffic analysis are **anonymous communications** which hide who talks with whom, but also other information such as frequency or duration.

These technologies are advantageous to criminals, but are also needed by many individuals that due to their jobs or roles have the need for internet privacy, even to protect their safety (e.g., journalists or abuse victims).

It is also important to remember that even if your job does not put you in danger, privacy is desirable for many reasons including protect yourself from intrusions and profiling.

Anonymous communications – Abstract model

IDs
Timing
Volume
Length
...

SENDERS

ANONYMOUS COMMUNICATION SYSTEM

RECEIVERS

Bitwise unlinkability
Use cryptography to make inputs and outputs to the anonymous communication systems appearance (bits) different

(re)packetizing + (re)schedule
Destroy patterns (traffic analysis resistance)

One-proxy problems
Low throughput
Corrupt Proxy or Proxy hacked / coerced
Real case: Penet.fi vs the church of scientology (1996)

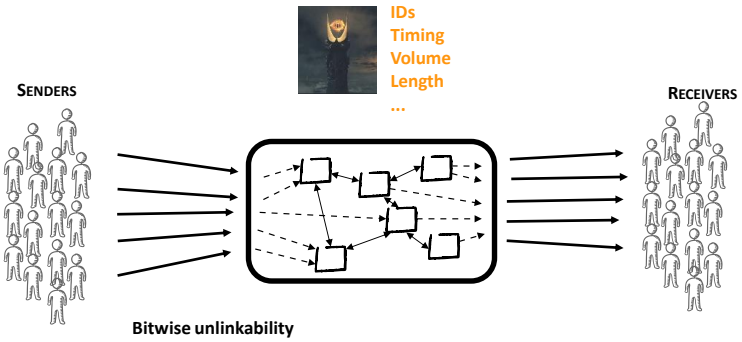
The basics of anonymous communications systems are to :

- *Provide bitwise unlinkability*: this means that the appearance of messages (i.e., the bits) when they come in the network and when they leave are different so that appearance itself cannot serve to trace messages. This is typically done by use of encryption.
- *Hide traffic patterns*: this means that anonymous communications try to change how traffic flows look like. The goal is that timing and sizes of packets cannot serve to trace communications. Typically, this is done by repacketizing packets (e.g., splitting messages in same-size packets) and re-scheduling packets (e.g., sending packets at regular intervals, or adding random delays to the packets).

These two properties are needed, but if they are enforced by a monolithic anonymous communication system (such as the one in the slide) this system:

1) Has limited throughput, i.e., it is unlikely that it can route simultaneous communications from millions of users).
2) The anonymous communication itself becomes a single point of failure for anonymity. If it is forced to reveal its logs users lose their anonymity. (https://www.spaink.net/cos/rnewman/anon/penet.html)

Anonymous communications – Abstract model



Bitwise unlinkability

Use cryptography to make inputs and outputs to the anonymous communication systems appearance (bits) different

(re)packetizing + (re)schedule + (re)routing

Destroy patterns (traffic analysis resistance)

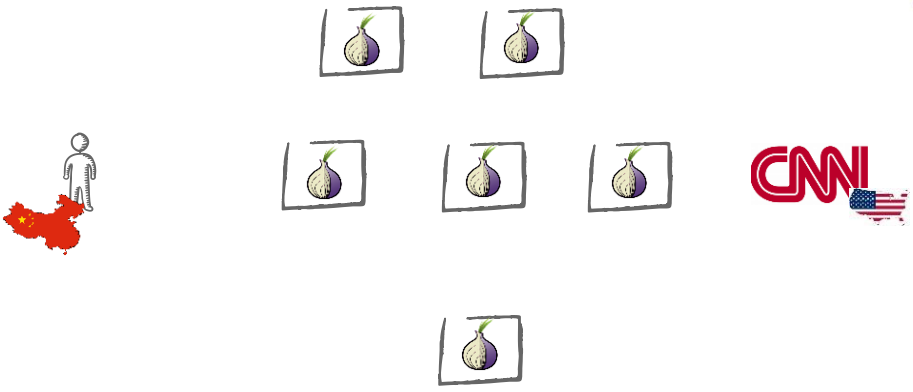
Load balancing

Distribute trust

Thus, modern anonymous communication systems rely on many nodes in different jurisdictions and messages are not only repacketized and rescheduled, but also rerouted.

This again implements separation of privilege and helps distributing trust.

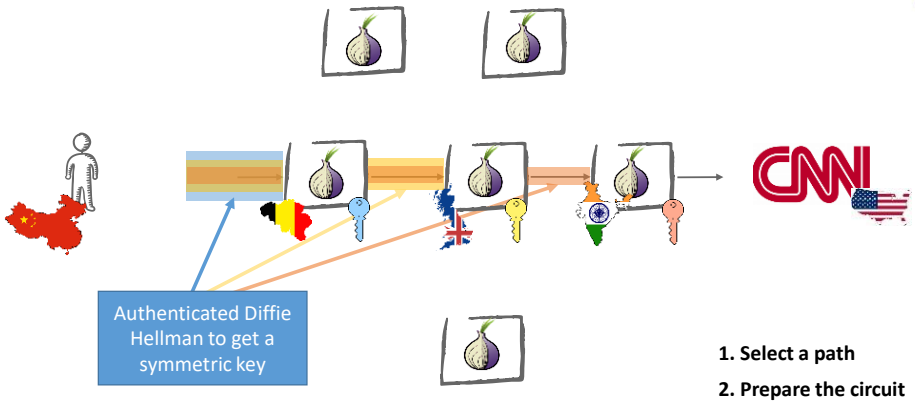
The Tor network – Onion routing



36

The main example of an active anonymous communication system is the Tor network (https://www.torproject.org/)

The Tor network – Onion routing



37

Tor uses onion encryption. It works as follows. When a user wants to communicate with a destination:

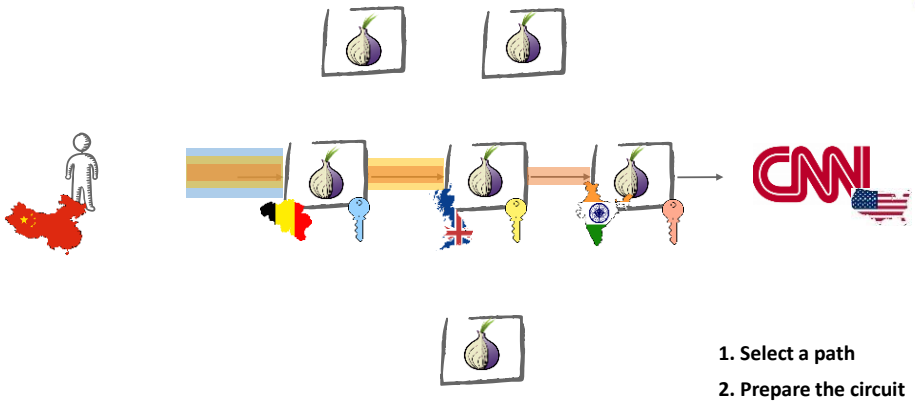
1) The user chooses a path (a series of nodes to route the traffic, currently 3 *onion routers*). These nodes' IP addresses and their public keys can be obtained from *Directory authorities* that maintain a list of all available Tor nodes at every point in time.
2) The user prepares the circuit, i.e., it agrees on a symmetric key with each of the nodes using an authenticated DH key agreement. Here, authenticated means that the onion routers sign their part of the key agreement protocol so that the user can be sure she is speaking with a Tor node.

   The key agreement is made directly with the first node (the *entry node*).

   With the second node (the *middle node*) the key is agreed using the entry node as intermediary. This way the middle node never sees who the user is.

   The key agreement is with the third node (the *exit node*) is made using the entry and middle nodes as intermediary. This way the exit node does not learn who the user nor the entry node are.

The Tor network – Onion routing



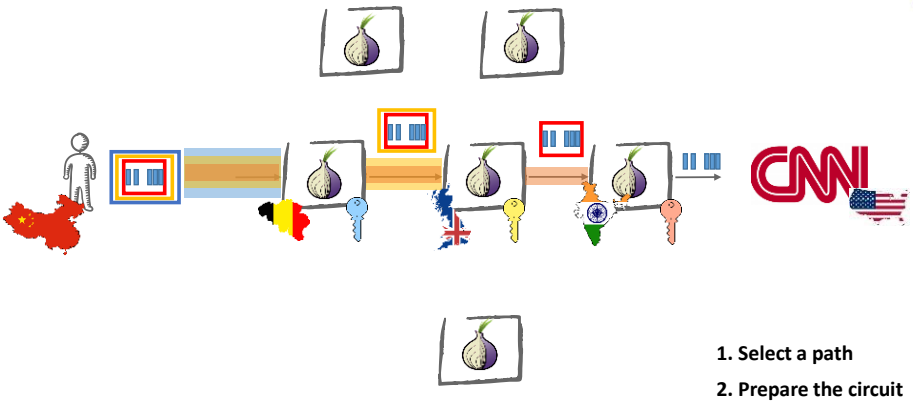
1. Select a path
2. Prepare the circuit
3. Send stream

38

3) Once the circuit is prepared the user can send messages. To send a message one encrypts it with the key of the exit node, then with the key of the middle node, and then with the key of the entry node. As the message advances in the network each node decrypts a layer (i.e., removes one “onion” layer) in such a way that messages have different appear when they enter and leave the network.

To respond, the exit node encrypts with the inverse onion: the inner layer is encrypted for the entry node, and then the middle.

The Tor network – Onion routing

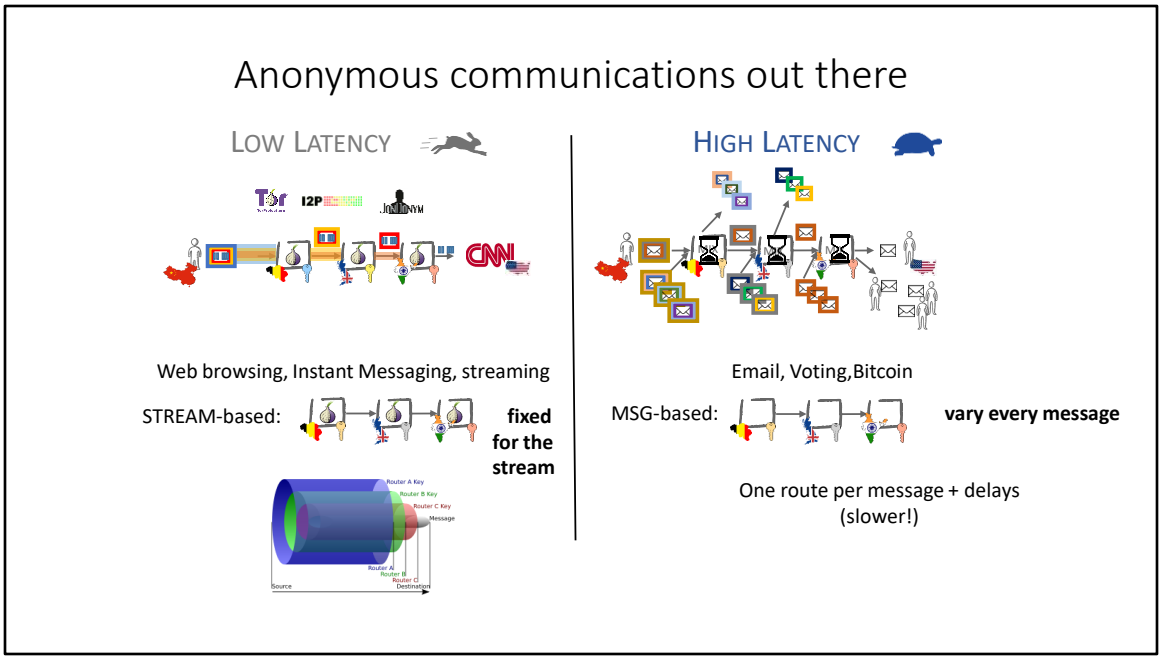


1. Select a path
2. Prepare the circuit
3. Send stream

39

3) Once the circuit is prepared the user can send messages. To send a message one encrypts it with the key of the exit node, then with the key of the middle node, and then with the key of the entry node. As the message advances in the network each node decrypts a layer (i.e., removes one “onion” layer) in such a way that messages have different appear when they enter and leave the network.

To respond, the exit node encrypts with the inverse onion: the inner layer is encrypted for the entry node, and then the middle.



Tor is the most widely example of a **low-latency anonymous communication network**. Low latency means that as messages come into a node they are decrypted and relayed to the next step without any delay other than the introduced by the processing of the packet in the node (decryption, queueing, etc).

Low-latency anonymous communications are typically used for browsing and streaming, and also for instant messaging applications in which users cannot tolerate delays beyond seconds.

One important characteristic of low-latency anonymous communications is that they are stream oriented. The user builds the circuit on the beginning relying on expensive public key operations, and obtaining symmetric keys. After the circuit is ready, all messages are encrypted with the symmetric keys to minimize the performance impact.

Another type of anonymous communication systems are **high-latency**, such as mix networks. In the simplest of these networks, each node (known as mix) waits until it receives a pre-defined number of messages (called the threshold). When these number is reached, the mix changes the appearance of the messages through

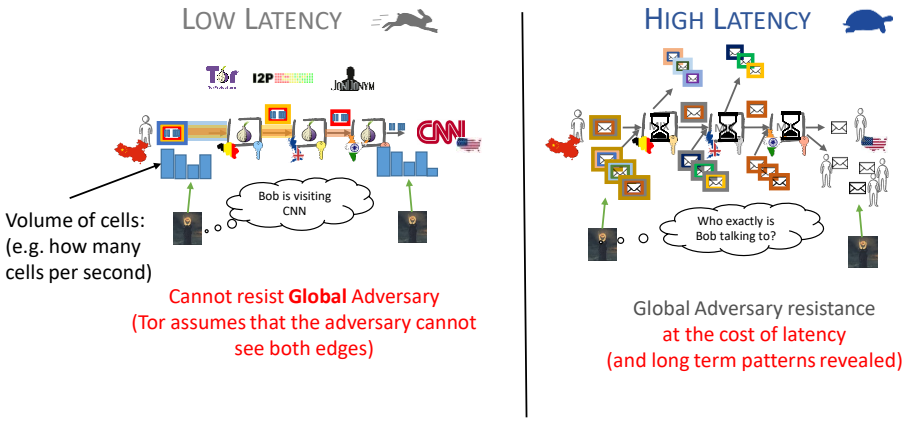
decryption and flushes all of them at the same time to the next mix or to the messages' destination.

High-latency anonymous communications are typically used for emailing and voting, as these do not require real time delivery of messages. More recently, they are also used in Bitcoin and other cryptocurrencies to enhance the privacy of transactions.

As opposed to low-latency anonymous communications, in high latency communications every message follows their own route. The user picks three nodes for every message. Encrypts this message with the public key of these nodes in an onion fashion and sends them. If messages are too long for public key encryption, each encryption is hybrid. Note that in this case *every* message requires public key operations, not like in a stream when only circuit building requires such operation.

We study more about high-latency anonymous communication networks in the advanced privacy technologies course (CS-523).

Anonymous communications out there



The performance gain of low-latency anonymous communications systems comes at the cost of only providing anonymity in presence of a *weak* adversary that cannot observe both edges of the network. Otherwise, due to the lack of delays inserted in the network, traffic patterns are preserved and an adversary can trace flows through the network.

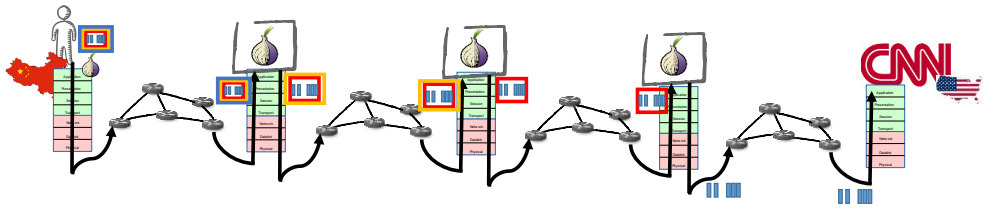
Mix-based communications, on the contrary, break traffic patterns by introducing latency. This enables them to protect against a global adversary, as there is no direct relation between the messages coming in the network and the messages going out. Even though the adversary cannot directly link incoming and outgoing messages, long term observation of the network can still reveal patterns of communications based on online/offline status of users. If every time Alice inputs a message in the network, Bob receives a message, at some point the adversary can infer with statistical significance that Alice and Bob are talking with each other.

Anonymous communication networks are overlay networks

Nodes in anonymous communication networks (e.g., onion routers in Tor) are **not** internet routers. They work at the application layer!

(overlay network = a computer network that is built on top of another network)

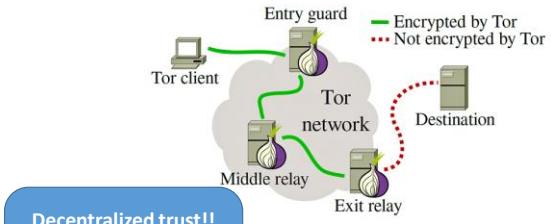
A more realistic view of how Tor traffic travels would be this



42

Note that when we talk about onion routers or mixes, these are not implemented at the network level. These routers operate at the application layer. In this sense, anonymous communication layers are **overlay networks**. They operate on top of another network (in this case the internet).

Anonymous communications vs. VPN



Different trust models!! Who is the adversary?

Centralized trust. No anonymity vs the VPN, or anyone seeing the VPN

Decentralized trust!! Provides privacy as long as the adversary cannot see both edges

https://www.privacyend.com/difference-between-tor-and-vpn/

Anonymous communication networks provide a much stronger protection than a VPN. In particular, anonymous communication networks follow the separation of privilege principle, such that none of the nodes in a path on its own can breach the anonymity of the user.

In a VPN, the VPN provider can be seen as unique node managing communications. If this node gets compromised, there is no anonymity guarantee for the users (recall the penet.fi case in slide 27).

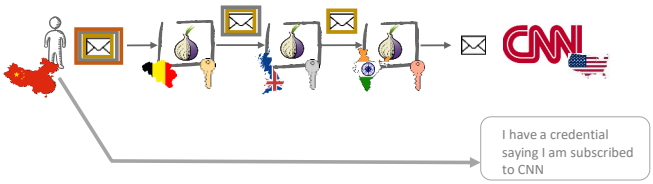
Anonymous communications at network layer what about the application layer?

me@cnn.com

me@cnn.com
logged in

Even if one uses anonymous communications, anonymity may be compromised when services require authentication.

Anonymous communications at network layer what about the application layer?



Anonymous credentials
Attribute-based credentials

When shown the server **cannot**
- Identify Alice (if her name is not provided)
- Learn anything beyond the info she gives (and what can be inferred)
- Distinguish two users with the same attributes
- Link multiple uses of the same credentials

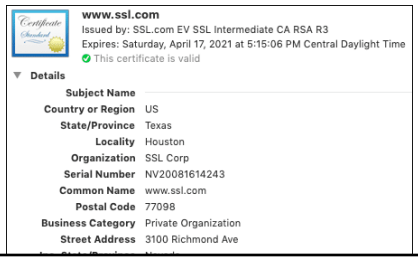
Anonymous credentials, also known as **Attribute-based credentials**, enable users to authenticate anonymously. What this means is that, when authentication and authorization do not depend on identity but they depend on an attribute (e.g., "being subscribed to", "being older than 18", "living in Lausanne"), one can prove this attribute without revealing one's identity.

Anonymous credentials, besides hiding the user identity:
- Only reveal the information that is proven (e.g., if the proven attribute is "being older than 18" the server cannot learn anything about the date of birth from the credential)
- Make two users with the same attribute undistinguishable: given two users subscribed to cnn.com, the service provider cannot distinguish them
- Link two credential shows from the same users: it is not possible to distinguish a user visiting cnn.com twice, from two users visiting cnn.com

Public Key Infrastructure
(usual internet authentication)

Signed by a trusted issuer
Certification of attributes
Authentication (secret key)

No data minimization
Users are identifiable
Users can be tracked
(Signature linkable to other contexts where PK is used)



Attribute based credentials

Signed by a trusted issuer
Certification of attributes
Authentication (secret key)

Data minimization
Users are anonymous
Users are unlinkable across contexts

Attribute-based credentials and Public Key infrastructure are similar in that:

- In both systems a trusted issuer does check that users actually have attributes and provide a signature on those attributes
- Require the user to have a secret in order to prove that she has some attributes.

But are also different:

- Attribute-based credentials minimize the data given to the service provider (only proof of possession of an attribute, not the value itself)
- In attribute-based credentials users cannot be identified and therefore non-tracked. PKI certificates, on the contrary contain names and always look the same. Therefore they enable tracking.

Other PETs examples

Private set intersection

a client and a server jointly compute the intersection of their private input sets in a manner that at the end the client learns the intersection and the server learns nothing (one-way PSI) or both learn the intersection (mutual PSI) -- private search

Blind Signatures

a server signs a message produced by a client without learning the content of the message -- eCash

Multiparty computation

parties to jointly compute a function over their inputs while keeping those inputs private –- compute total computations (statistics)

Private information retrieval

cryptographic method that allows a user to query a database without the server knowing which item the user asked for

Privacy Quantification

No Free Lunch Theorem [1]:

For every algorithm that outputs a D with even a sliver of utility, there is some adversary with a prior such that privacy is not guaranteed

[1] "No Free Lunch in Data Privacy", Kifer and Machanavajjhala, 2011