cybersecurity centre
GCHQ Academic Centre of Excellence EPSRC

Cybersecurity Southampton

CyberSecuritySoton.org [w]

@CybSecSoton [fb & tw]

UNIVERSITY OF Southampton

COMP6224
Privacy, Data Anonymisation and Anonymous Network

Dr Federico Lombardi

f.lombardi@soton.ac.uk

Lecture Outline

- What is privacy
- Privacy threats
- Privacy enhancing solutions (PETs)
- Data Anonymisation
- Anonymous Network and TOR

Learning Outcomes

- At the end of this lecture you should be able to:
  - Provide a definition of privacy
  - Provide examples of privacy threats
  - Link privacy enhancing technologies to privacy threats
  - Understand main approach for data anonymisation
  - Understand anonymous network and TOR

The Dead of Privacy

facebook

Where privacy dies.

What is Privacy?

Privacy is a concept in disarray. Nobody can articulate what it means. As one commentator has observed, privacy suffers from "an embarrassment of meanings."

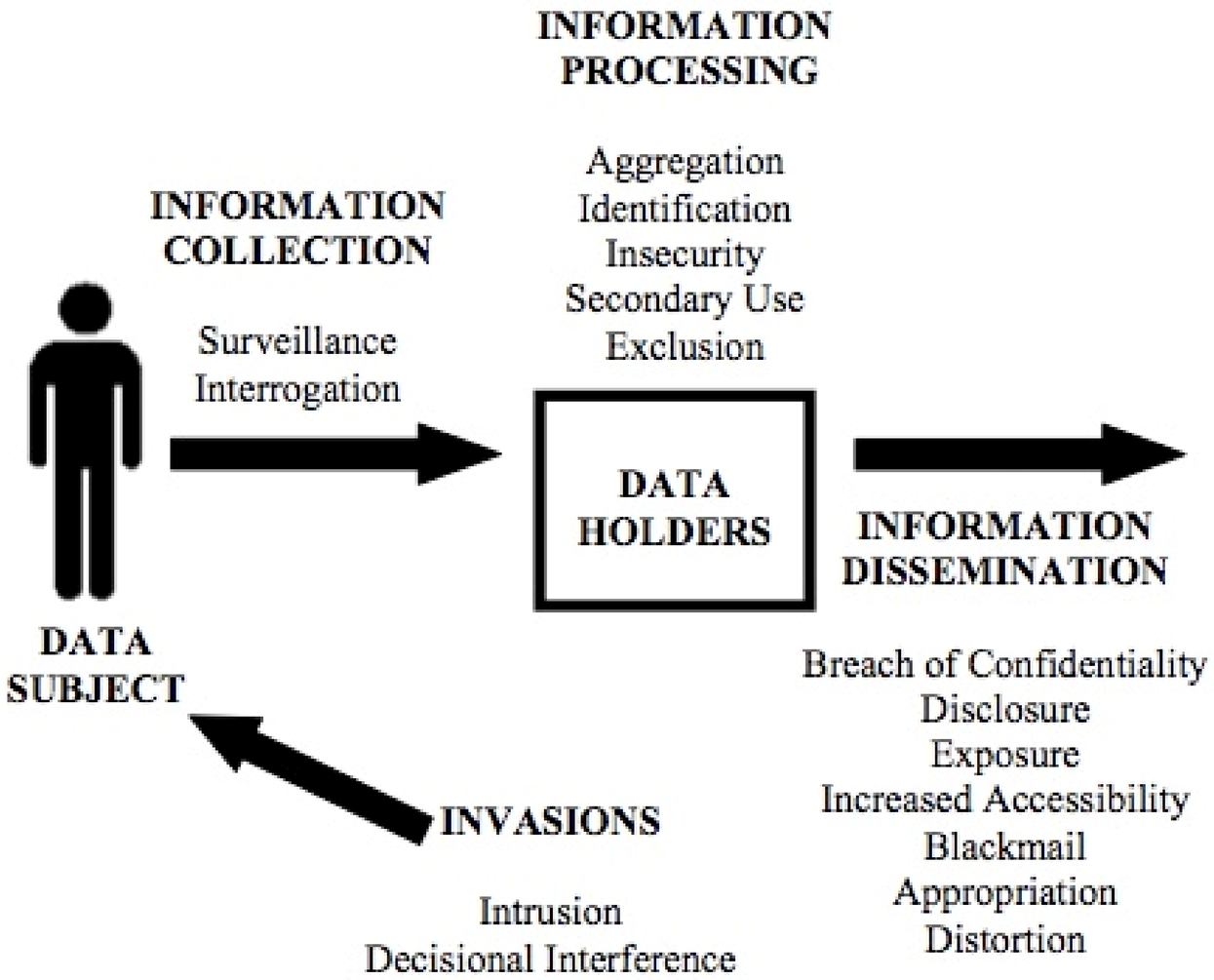
Privacy Definitions

"the right to be let alone"
Warren and Brandeis (1890)

"the right of the individual to decide what information about himself should be communicated to others and under what circumstances" *Westin (1970)*

"the freedom from unreasonable constraints on the constructions of one's identity" *Agre & Rotenberg (2001)*

Solove's Privacy Taxonomy



Solove Privacy Taxonomy in Details

INFORMATION PROCESSING

Aggregation
Combining of various pieces of personal information
A credit bureau combining an individual's payment history from multiple creditors.

Insecurity
Carelessness in protecting information from leaks or improper access
An ecommerce website allowing others to view an individual's purchase history by changing the URL. (e.g. enterprivacy.com?id=123)

Identification
linking of information to a particular individual
A researcher linking medical files to the Governor of a state using only date of birth, zip code and gender.

Secondary Use
Using personal information for a purpose other than the purpose for which is was collected
The U.S. Government using census data collected for the puprose of apportioning Congressional districts to identify and intern those of Japanese descent in WWII.

Exclusion
Failing to let an individual know about the information that others have about her and participate in its handling or use
A company using customer call history, without the customer's knowledge, to shift their order in a queue (i.e. "Your call will be answer in the order [NOT] received")

INFORMATION COLLECTION

Surveillance
Watching, listening to, or recording of an individual's activities
A website monitoring cursor movements of a visitor while visiting the website.

Interrogation
Questioning or probing for personal information
An interviewer asking an inappropriate question, such as marital status, during a employment interview.

INVASION

Intrusion
Disturbing an individual's tranquility or solitude
An augmented reality game directing players onto private residential property.

Decisional Interference
Intruding into an individual's decision regarding her private affairs
A payment processor declining transactions for contraceptives.

Based on Dan Solove's A Taxonomy of Privacy

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=667622

Information Processing
Aggregation
Insecurity
Identification
Secondary Use
Exclusion

Information Collection
Surveillance
Interrogation

Invasion
Intrusion
Decisional Interference

Information Dissemination
Breach of Confidentiality
Disclosure
Exposure
Increased Accessibility
Blackmail
Appropriation
Distortion

INFORMATION DISSEMINATION

Breach of Confidentiality
Breaking a promise to keep a person's information confidential
A doctor revealing patient information to friends on a social media website.

Disclosure
Revealing truthful information about a person that impacts her security or the way others judge her character
A government agency revealing an individual's address to a stalker, resulting in the individual's murder.

Exposure
Revealing an individual's nudity, grief, or bodily functions
A store forcing a customer to remove clothing revealing a colostomy bag.

Increased Accessibility
Amplifying the accessibility of personal information
A court making proceeding searchable on the Internet without redacting personal information.

Blackmail
Threatening to disclose personal information
A dating service for adulters charging customers to delete their accounts.

Appropriation
Using an individual's identity to serve the aims and interests of another
A social media site using customer's images in advertising

Distortion
Disseminating false or misleading information about an individual
A creditor reporting a paid bill as unpaid to a credit bureau.

Information Collection

- ***Surveillance***
  - is the watching, listening to, or recording of an individual's activities.
- ***Interrogation***
  - consists of various forms of questioning or probing for information.

Information Collection: Surveillance

- *"Smart meters are presented as an environmental and power-saving initiative. But it's a highly surveillant model. It can tell how many showers you have had, when you are cooking, when you are in and out of the home."*

- *"We take energy consumption data from smart meters and sensors. We analyse it and build a highly personalised profile for each and every utility customer"* (0nzo, British Analytics Company).

Information Collection: Surveillance

- ***Surveillance***
  - is the watching, listening to, or recording of an individual's activities.

- Examples:
  - IoT
  - Voice Assistants

BRIAN BARRETT SECURITY 02.07.17 08:03 PM

HOW TO STOP YOUR SMART TV FROM SPYING ON YOU

THIS WEEK, VIZIO, which makes popular, high-quality, affordable TV sets, agreed to pay a $2.2 million fine to the FTC. As it turns out, those same TVs were also busily tracking what their owners were watching, and shuttling that data back to the company's servers, where it would be sold to eager advertisers.

Information Collection: Surveillance

- ***Surveillance***
  - is the watching, listening to, or recording of an individual's activities.

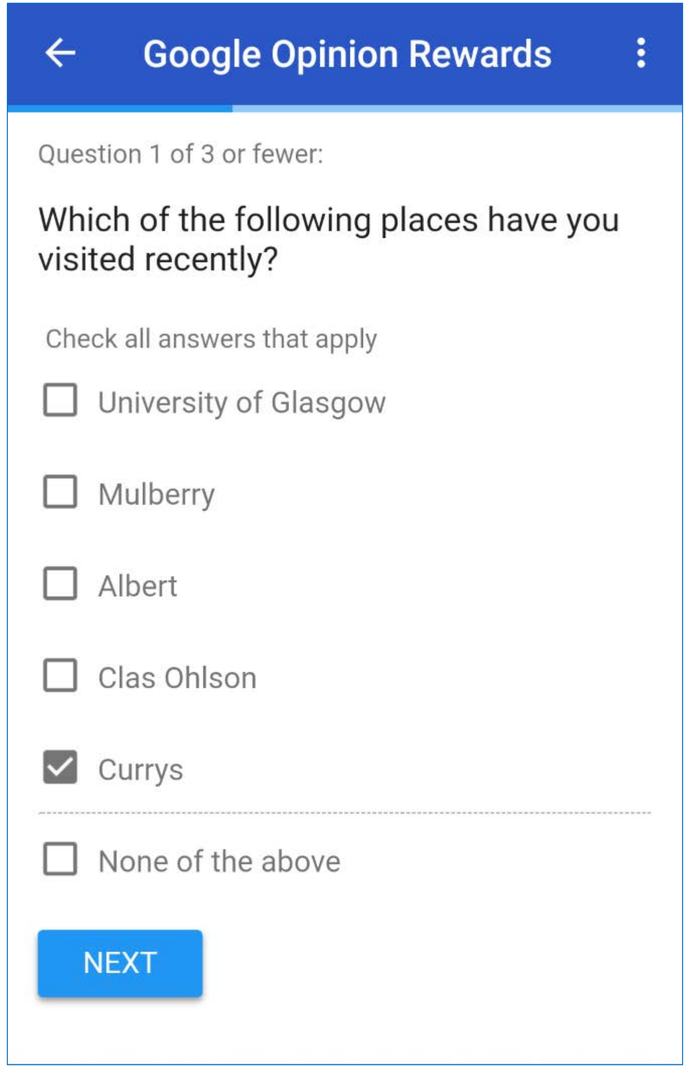
- Other examples: malicious smartphone apps

Angry Birds and 'leaky' phone apps targeted by NSA and GCHQ for user data

- US and UK spy agencies piggyback on commercial data
- Details can include age, location and sexual orientation
- Documents also reveal targeted tools against individual phones

Information Collection: Inerrogation

- ***Interrogation***
  - consists of various forms of questioning or probing for information.



Information Processing

- ***Aggregation***
  - involves the combination of various pieces of data about a person.
- ***Identification***
  - linking information to particular individuals.
- ***Insecurity***
  - involves carelessness in protecting stored information from leaks and improper access.
- ***Secondary use***
  - is the use of information collected for one purpose for a different purpose without the data subject's consent.
- ***Exclusion***
  - concerns the failure to allow the data subject to know about the data that others have about her and participate in its handling and use.

Information Processing: Aggregation

- ***Aggregation***
  - involves the combination of various pieces of data about a person.

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF
Welcome to The Not-So Private Parts where technology & privacy collide
FULL BIO

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT +2.2%, for example, has figured out how to

- ***Identification***
  - linking information to particular individuals.

On the Web's Cutting Edge, Anonymity in Name Only

Posted on August 3, 2010 by juliaangwin — No Comments ↓

The Wall Street Journal, Page One

You may not know a company called [x+1] Inc., but it may well know a lot about you.

From a single click on a web site, [x+1] correctly identified Carrie Isaac as a young Colorado Springs parent who lives on about $50,000 a year, shops at Wal-Mart and rents kids' videos. The company deduced that Paul Boulifard, a Nashville architect, is childless, likes to travel and buys used cars. And [x+1] determined that Thomas Burney, a Colorado building contractor, is a skier with a college degree and looks like he has good credit.

The company didn't get every detail correct. But its ability to make snap assessments of individuals is accurate enough that Capital One Financial Corp. uses [x+1]'s calculations to instantly decide which credit cards to show first-time visitors to its website.

Information Dissemination

- ***Breach of confidentiality***
  - is breaking a promise to keep a person's information confidential.
- ***Disclosure***
  - involves the revelation of truthful information about a person that impacts the way others judge her character.
- ***Exposure***
  - involves revealing another's nudity, grief, or bodily functions.
- ***Increased accessibility***
  - is amplifying the accessibility of information
- ***Blackmail***
  - is the threat to disclose personal information
- ***Appropriation***
  - involves the use of the data subject's identity to serve the aims and interests of another.
- ***Distortion***
  - consists of the dissemination of false or misleading information about individuals.

Information Dissemination: Breach to Confidentiality

- ***Breach of confidentiality***
  - is breaking a promise to keep a person's information confidential.

Equifax finds more victims of 2017 breach

1 March 2018 Share

US politicians have criticised Equifax, saying that it "botched" its earlier response to the breach

The massive data breach suffered by credit-rating company Equifax hit more people than previously thought, the company has reported.

Information Dissemination: Exposure

- ***Exposure***
  - involves revealing another's nudity, grief, or bodily functions.

Massive Leak of Celebrity Nude Photos Calls Cloud Security Into Question

PREVIOUS CONTRIBUTORS

SEP 1, 2014 | VULNERABILITY MANAGEMENT

Information Dissemination: Appropriation

- ***Appropriation***
  - involves the use of the data subject's identity to serve the aims and interests of another.

Identity fraud up by 57% as thieves 'hunt' on social media

5 July 2016 | UK

Share

The BBC's Angus Crawford went on the trail of identity fraudsters last year

Information Dissemination: Distortion

- ***Distortion***
  - consists of the dissemination of false or misleading information about individuals.

Shropshire internet troll left counting cost over £100,000 legal bill

Telford | News | Published: Mar 7, 2015

> An internet "troll" from Shropshire was today left counting the cost after being hit with a record £100,000 legal bill for using his home computer to target an American lawyer for groundless abuse.

Subscribe to our daily newsletter Email address: Sign Up

Invasions

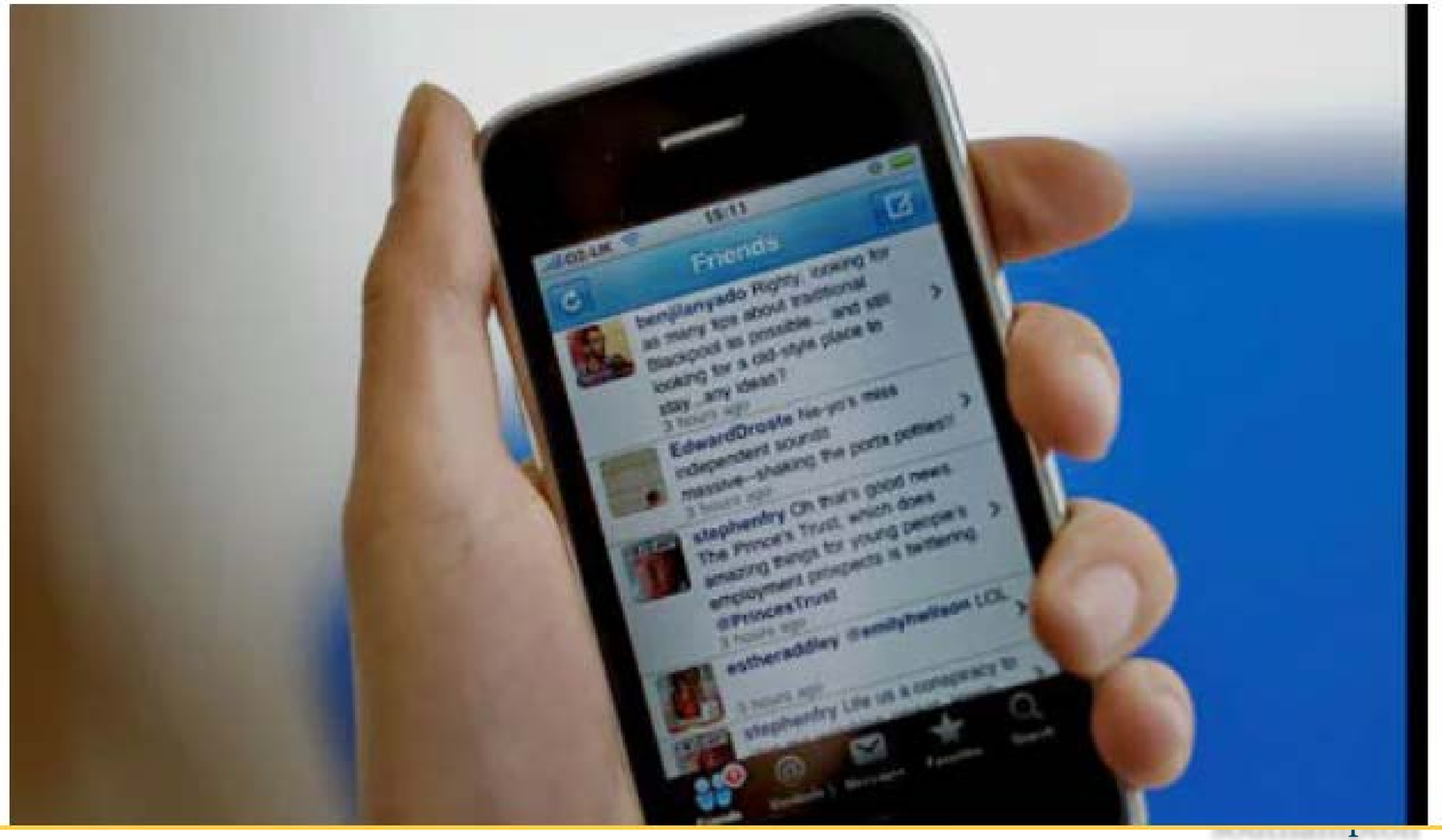
- ***Intrusion***
  - concerns invasive acts that disturb one's tranquility or solitude
- ***Decisional interference***
  - involves the government's incursion into the data subject's decisions regarding her private affairs

Invasions: Intrusion

- ***Intrusion***
  - concerns invasive acts that disturb one's tranquility or solitude

Social networking sites fuelling stalking, report warns

Smartphones and social networking sites are making it much easier for stalkers to target victims, say charities



Privacy Enhancing Technologies (PETS)

- Tools, mechanism or architectures that aim to mitigate privacy concerns
  - While allowing users to enjoy the benefits of modern technologies
- PETS can be applied to communications or to existing databases
- PETS can be deployed either by individual users or by organizations
- They can be divided in 3 main categories
  - Privacy as Confidentiality
  - Privacy as Control
  - Privacy as Practice

Privacy as Confidentiality

Privacy as Confidentiality

Data Anonymisation

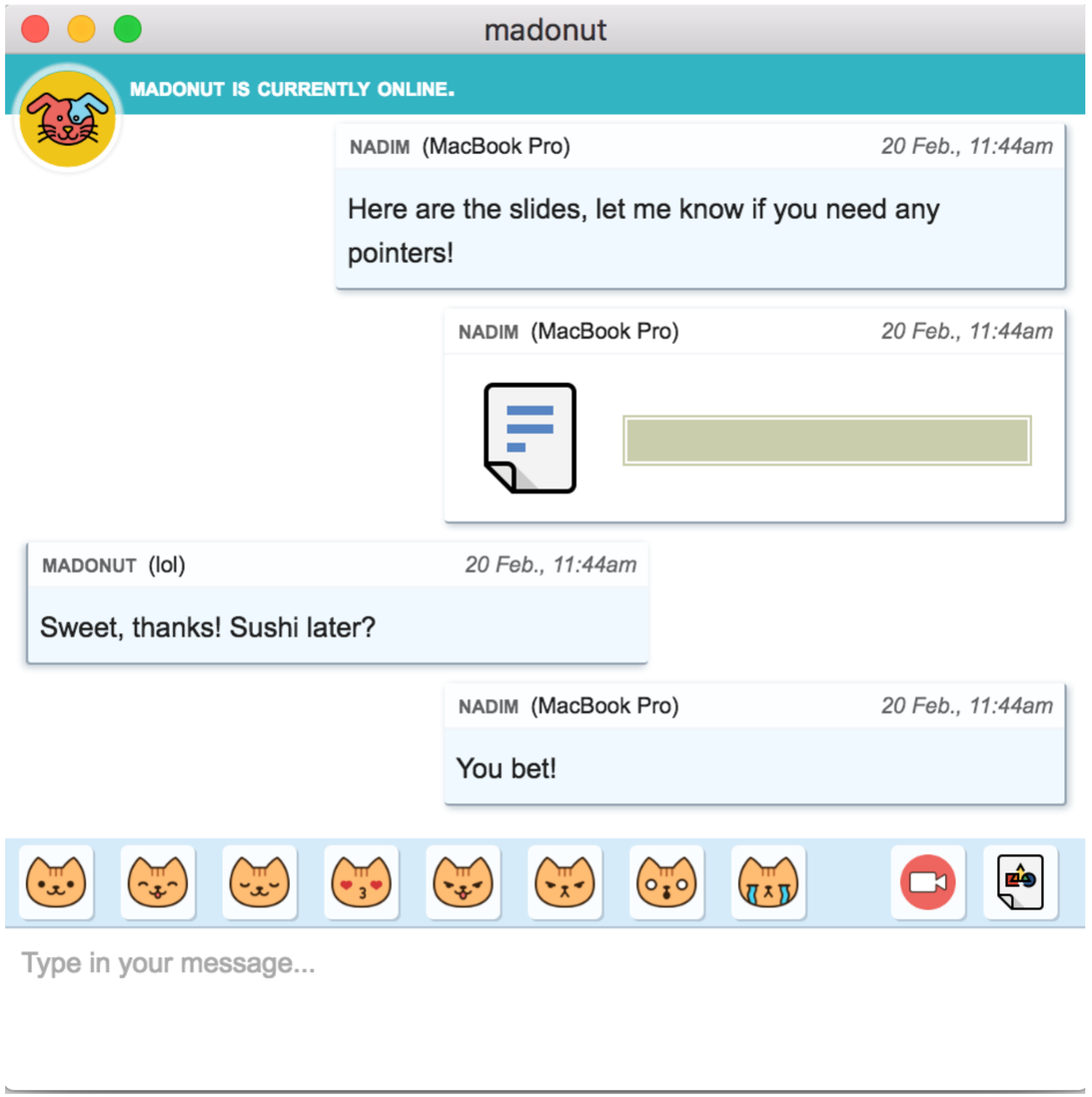
Secure messaging

Anonymous communications

“the right to be let alone”
Warren and Brandeis

Data minimization

An Example: Cryptocat



Privacy as Control

UNIVERSITY OF Southampton

Privacy as Control

Anonymous Credentials

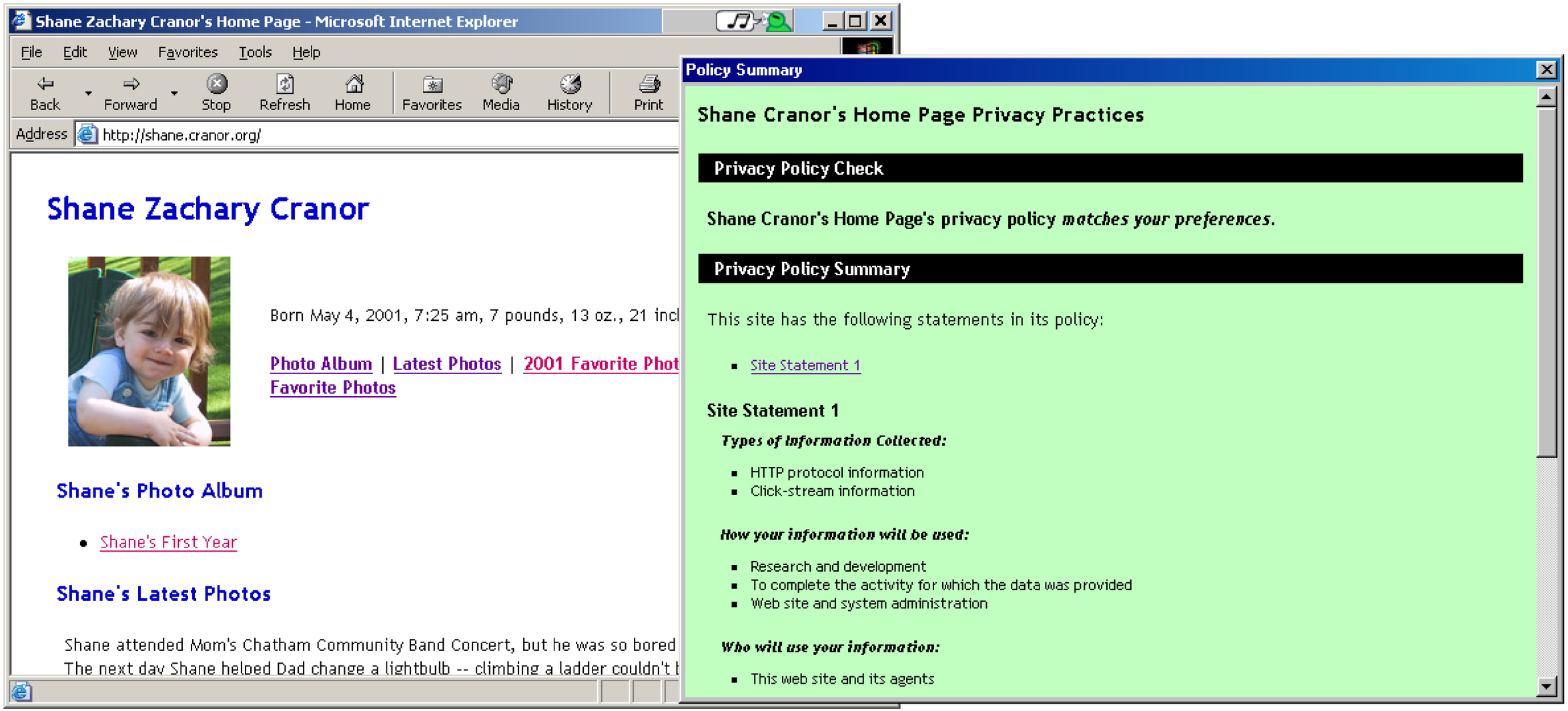
Privacy policy languages

Purpose based access control

“the right of the individual to decide what information about himself should be communicated to others and under what circumstances” Westin

Compliance

Privacy Bird



Privacy as Practice

Privacy as Practice

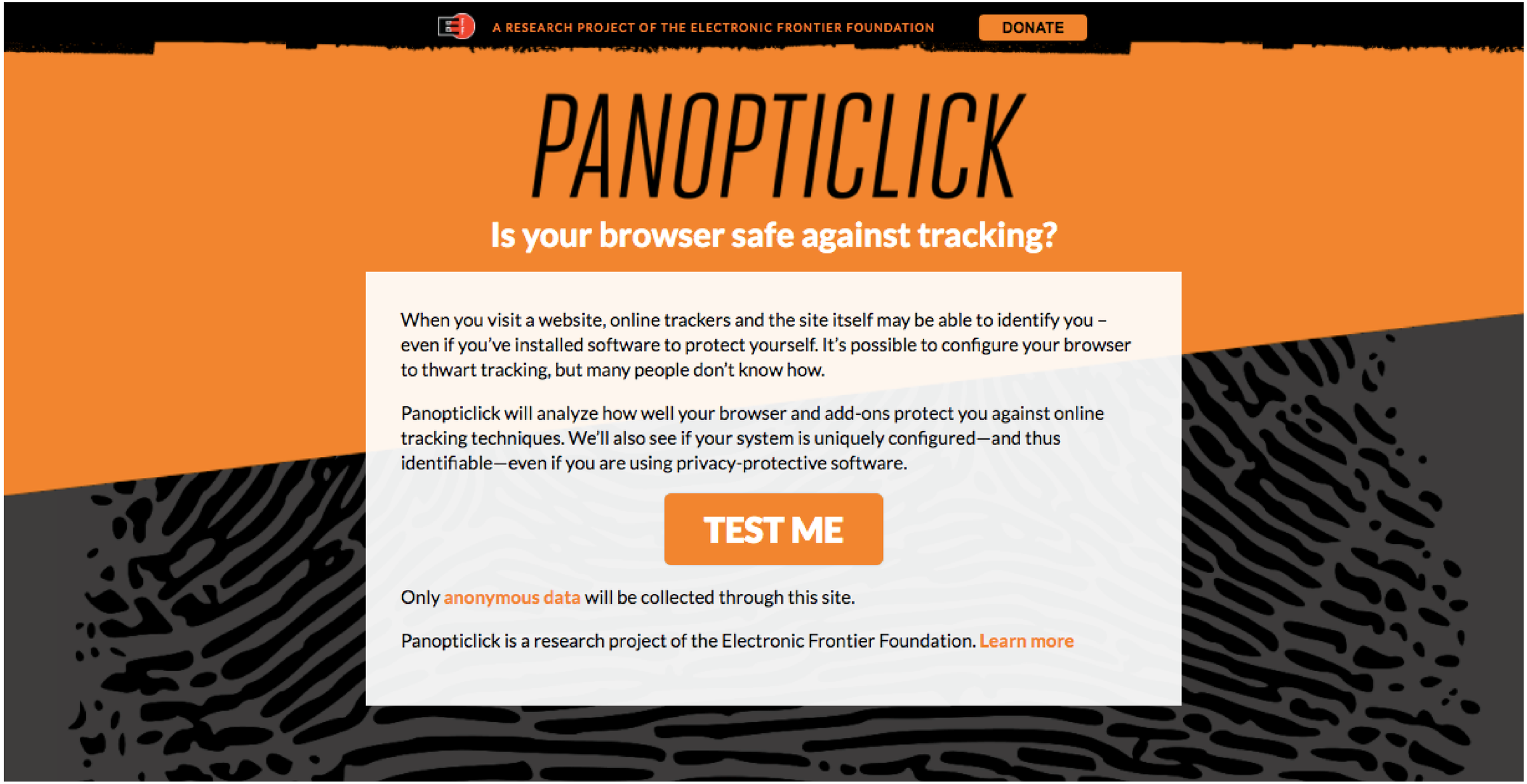
Feedback and awareness tools

Privacy nudges

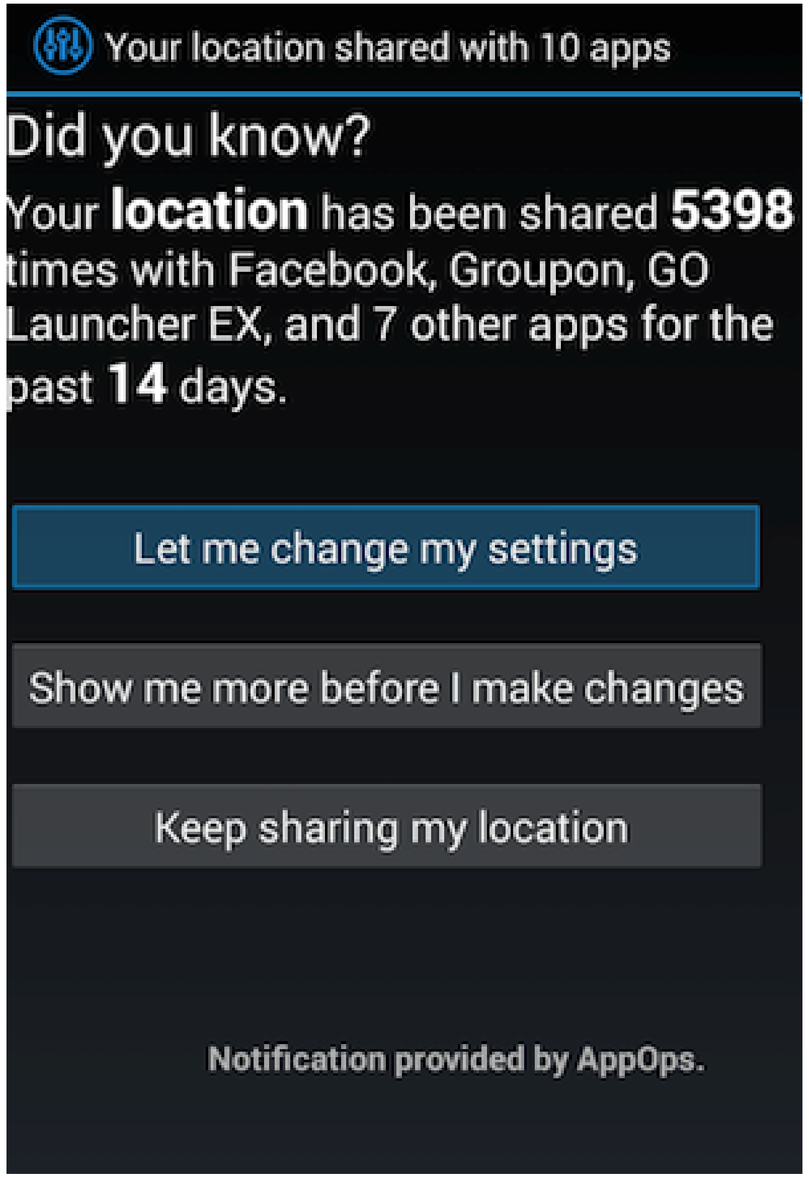
“the freedom from unreasonable constraints on the constructions of one’s identity” Agre

aid in privacy decision making

Web fingerprinting



Privacy Nudges



PET Families

- PET can be divided in two main families:

1. **Soft Privacy Technology**:
   - It assumes that the third-party can be trusted for the processing of data.
   - Model is based on compliance, consent, control and audit.
   - Example technologies are access control, tunnel encryption (SSL/TLS).

2. **Hard Privacy Technology**:
   - It assumes that third-parties cannot be trusted.
   - No single entity can violate the privacy of the user.
   - The data protection goal is data minimization and reduction of the trust in third-parties.
   - Examples of such technology is onion routing (TOR).

Summary

- Privacy is about protecting the association between an individual and his/her personal information
- There are three main privacy research paradigms
  - Privacy as confidentiality
  - Privacy as control
  - Privacy as practice

Recommended Readings

- Privacy Research Paradigms. Available for download from the module wikipage
- Daniel J. Solove. A Taxonomy of Privacy. Available at: https://www.law.upenn.edu/journals/lawreview/articles/volume154/issue3/Solove154U.Pa.L.Rev.477(2006).pdf

What about Data Anonymisation?

Data Anonymisation

- Why Data Anonymisation:
  - For privacy preserving data analytics

- Data Anonymisation techniques:
  - k-anonimity
  - l-diversity
  - t-closeness
  - differential privacy

Classification of Attributes

- **Explicit identifiers**
  - Identify a user
  - E.g name, lastname, passport number, etc.
- **Quasi-identifiers**
  - E.g Date of birth, Age, Zip code, phone number
- **Sensitive attributes**
  - E.g diseases, salaries, etc.

These attributes is what the researchers need, so they are always released directly

An Example

Key Attributes | **Quasi-identifiers** | **Sensitive attributes**

ID	Name	DOB	Gender	Zipcode	Disease
12345	Andre	1/21/76	Male	53715	Heart Disease
56789	Beth	4/13/86	Female	53715	Hepatitis
52131	Carol	2/28/76	Male	53703	Brochitis
85438	Dan	1/21/76	Male	53703	Broken Arm
91281	Ellen	4/13/86	Female	53706	Flu
11253	Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity

- A record has to be indistinguishable from at least k-1 other records with the respect to the quasi-identifiers.
- Each class of equivalence has to contain at least k records which have the same values for the quasi identifiers.

Original Database

Name	Zipcode	Age	Disease
Hilary	47677	29	Heart Disease
Jenny	47602	22	Heart Disease
Bob	47678	27	Heart Disease
Izzy	47905	43	Flu
John	47909	52	Heart Disease
Fred	47906	47	Cancer
Sam	47605	30	Heart Disease
Carl	47673	36	Cancer
Sarah	47607	32	Cancer

Released Database

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Achieving k-Anonymity

- **Generalization**
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- **Suppression**
  - When generalization causes too much information loss
    - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations
  - … usually without any clear notion of utility

Attacks against k- Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Umeko	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

L-diversity: intuition

[Machanavajjhala et al. ICDE ‘06]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be “diverse” within each quasi-identifier equivalence class

Distinct l-diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks

….	Disease
	HIV
	HIV
	HIV
	HIV
	HIV
	HIV
	HIV
	Bronchitis
	Pneumonia

8 records have HIV

2 records have other values

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	5K	Gastritis
476**	2*	9K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

t-closeness

[Li et al. ICDE ‘07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

Differential Privacy: The Problem to Address

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

How many men have the disease ? 2

What is the average age / weight of men who have the disease ? 50 / 101

insertion of a new record

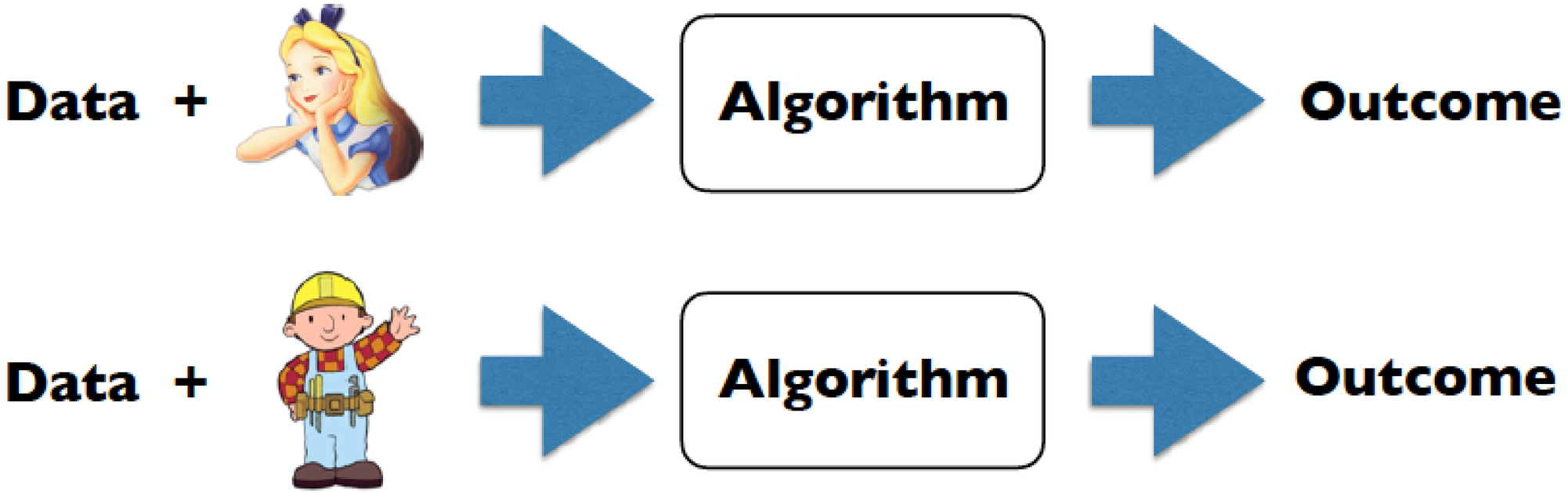
Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

We can deduce the exact age / weight of the new record

Differential privacy: Intuition



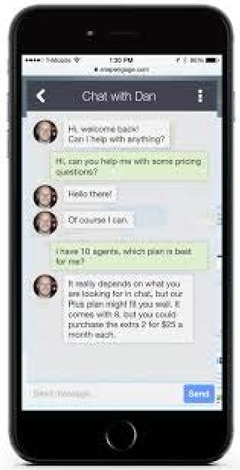
Participation of a person does not change the outcome

UNIVERSITY OF Southampton

Differential Privacy Algorithm

An algorithm is **differentially private** if an observer seeing its output cannot tell if a particular individual's information was used in the computation.

Google



Recommended Readings

- Static Data Anonymization Part I: Multidimensional Data. Available at https://secure.ecs.soton.ac.uk/noteswiki/w/File:L05-Anonymization.pdf
- l-diversity: Privacy Beyond k-Anonymity: Available at: https://dl.acm.org/citation.cfm?id=1217302
- t-closeness: Privacy Beyond K-anonymity and l-diversity. Available at http://ieeexplore.ieee.org/document/4221659/
- Algorithmic foundations of differential privacy. Available at: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

What about Anonymous Network?

Simple Proxying

We could proxy through P

A → P → B

- But P still knows about A and B

The simple ”school filter” bypass technique

Plenty of open proxy servers online

- But P will know A and B communicated
- P will know what A and B sent to each other
- We still know that A connected to P
- And B knows that P connected to it

Open Proxy Servers

Public Proxy Servers

Proxy Servers - Page 1 of 6 — **Ugrade to VPN!**

Home | Add Proxy | Useful Links

Proxy List | ↑ Domain | ↓ Rating | ↑ Country | ↓ Access Time | ↓ Uptime | ↓ Online Time | ↑ Last Test

HideMyAss VPN offers you protection and peace of mind for your personal data whenever you surf the internet - wherever you are.

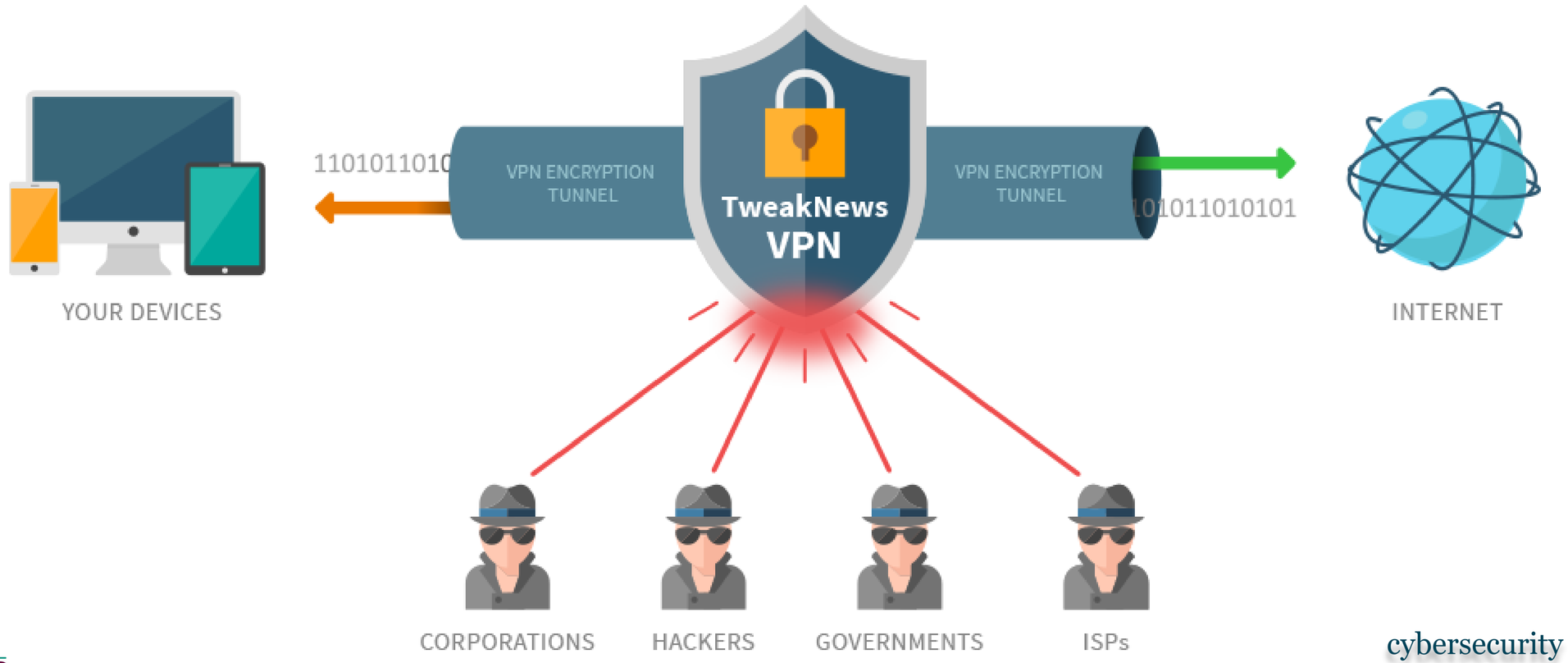
Domain	Rating	Country	Access Time	Uptime %	Online Since	Last Test	Features	
hispaproxy.eu	70	Germany	1.0	87	44 minutes	45 minutes	HiAn	SSL
popunderfreeproxy.eu	81	Netherlands	1.0	98	14 hours	0 seconds	HiAn	SSL
europroxy.pw	81	France	1.0	98	2 days	30 minutes	HiAn	SSL
freesite.work	88	United States	0.4	98	5 hours	15 minutes	HiAn	SSL
surfingonmyown.com	26	United States	0.7	38	-	15 minutes	HiAn	SSL
proxyeuro.pw	82	France	0.9	97	1 day	15 minutes	HiAn	SSL
time2hide.one	85	United States	0.4	99	3 hours	44 minutes	HiAn	SSL
brokenleg.cf	75	Netherlands	1.4	98	4 days	15 minutes	HiAn	SSL

Simple Proxying

- Why is this a problem?
  - Direct connection
    - Between your machine and the proxy
    - Between the proxy and the target
  - If P is malicious, they have everything you sent and received
  - You have to allow P to MITM – do you trust it?
  - What if P was compromised/hacked?
  - What if P was later seized by the authorities?
  - What if P is being run by the authorities?

Alternative? VPN connection

VPN is a Virtual Private Network that establish a connection towards a (trusted) server and all your internet traffic is encrypted in this channel



Why VPN are useful?

- Used to access sensitive service or data in a company from outside
- Used to anonymise the traffic as the ISP will no longer know which websites you surf as it will only see a connection towards the VPN server
- Used to simulate your current position to the one of the VPN server
  - Useful for banned website in specific Countries
  - Useful to access media content of another Country (like pay-per-view TV, etc…)

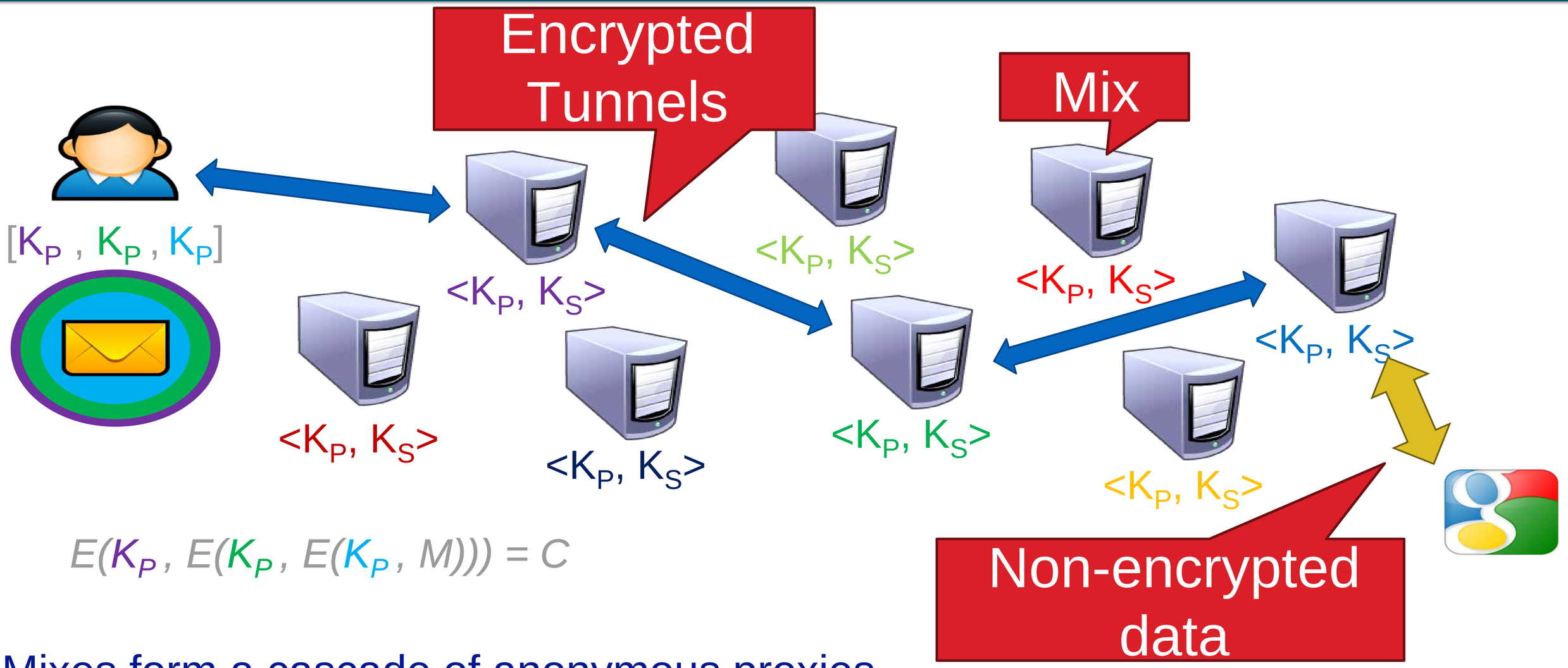
Mix Networks

- A different approach to anonymity

- Originally designed for anonymous email
  - David Chaum, 1981
  - Concept has since been generalized for TCP traffic

- Hugely influential ideas
  - Onion routing
  - Traffic mixing
  - Dummy traffic (a.k.a. cover traffic)

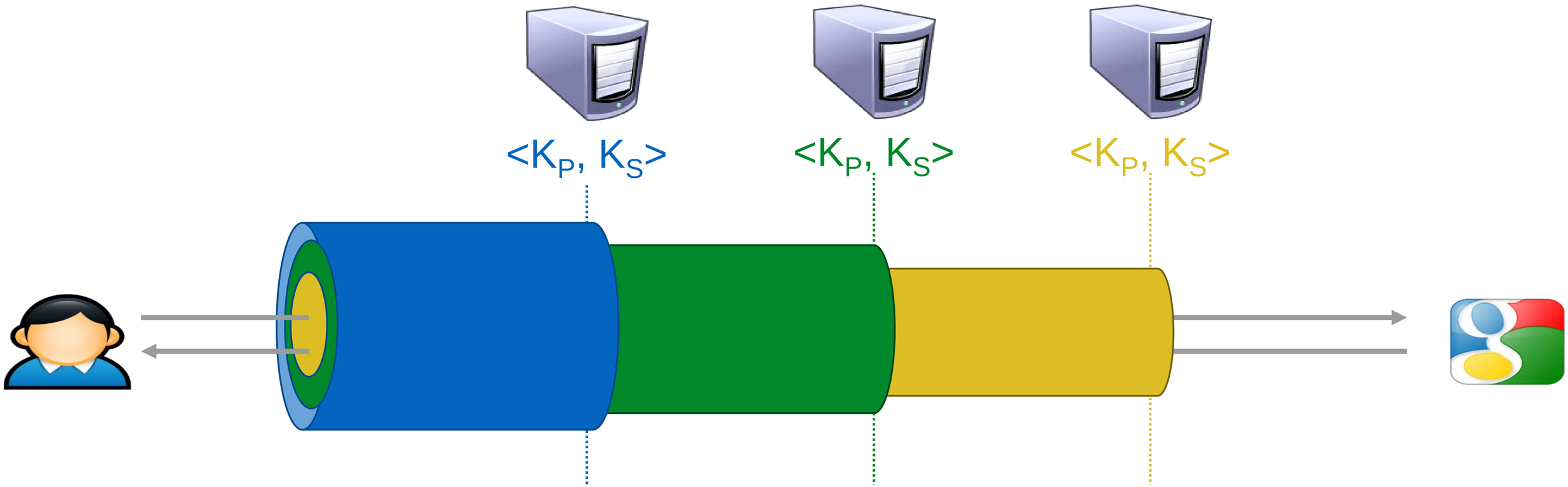
Mix Proxies and Onion Routing



Mixes form a cascade of anonymous proxies

All traffic is protected with layers of encryption

Another View of Encrypted Paths

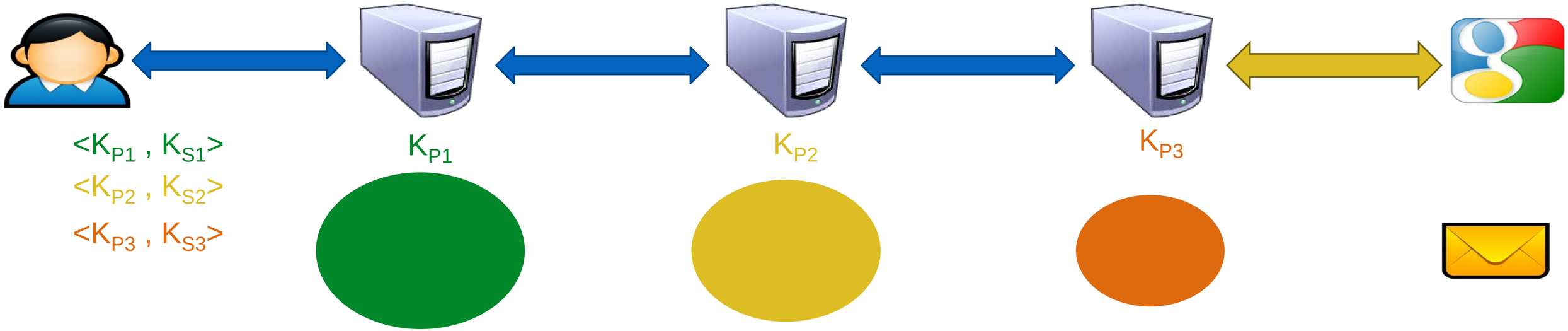


Return Traffic

In a mix network, how can the destination respond to the sender?

During path establishment, the sender places keys at each mix along the path

- Data is re-encrypted as it travels the reverse path



Tor: The 2nd Generation Onion Router

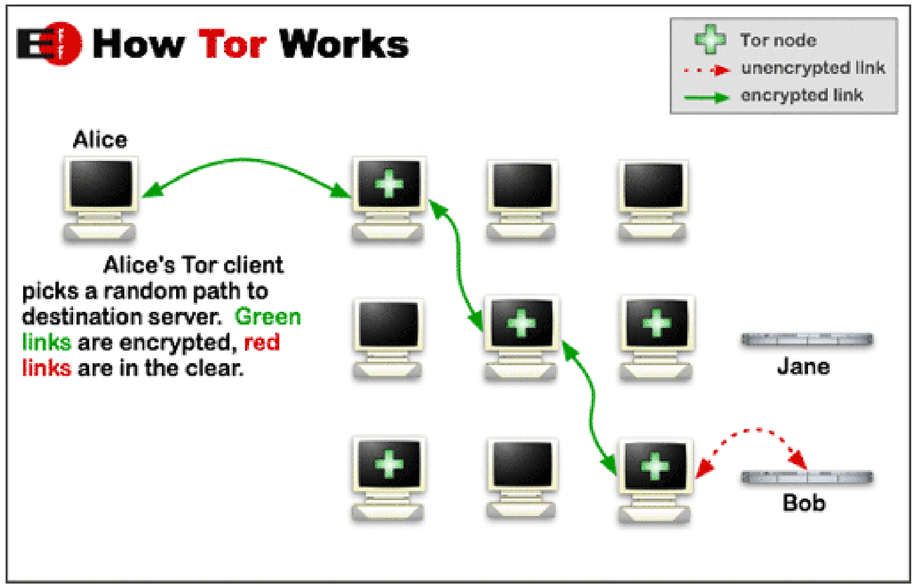
- Basic design: a mix network with improvements
  - Takes bandwidth into account when selecting relays
    - Mixes in Tor are called relays
  - Introduces hidden services
    - Servers that are only accessible via the Tor overlay

Guard Nodes & Exit Nodes

Traffic passed through 3+ servers

Guard Nodes: the input nodes (can know the identity of the sender)

Exit Nodes: can see your traffic if it is unencrypted



Hidden Services

- Tor is very good at hiding the source of traffic
  - But the destination is often an exposed website
- What if we want to run an anonymous service?
  - i.e. a website, where nobody knows the IP address?
- Tor supports Hidden Services
  - Allows you to run a server and have people connect
  - … without disclosing the IP or DNS name
- Many hidden services
  - Tor Mail, Tor Char
  - DuckDuckGo
  - Wikileaks

Reccommended Readings

- G. Danezis, C. Diaz. A survey of Anonymous Communication Channels.
- Tor Overview. Available at https://www.torproject.org/about/overview.html.en
- Tor Onion Services aka Hidden Services. Available at https://www.torproject.org/docs/onion-services.html.en
- Tor: The Second Generation Onion Router. Available at: https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf