Web Tracking

Dr. Chen Zhang
Department of Computer Science
The Hang Seng University of Hong Kong

Slides credit in part from A. Juels, N. Bielova, M. Langheinrich, S. Yuan, H. Fosdick, A. Judmayer, V. Shmatikov, I. Gonshorovitz, and L. Cranor

It's the Internet! Of course they know you're a dog. They also know your favorite brand of pet food and the name of the cute poodle at the park that you have a crush on!

- **Billions of web users surf the web daily**
- **Some websites have billions of user accounts**

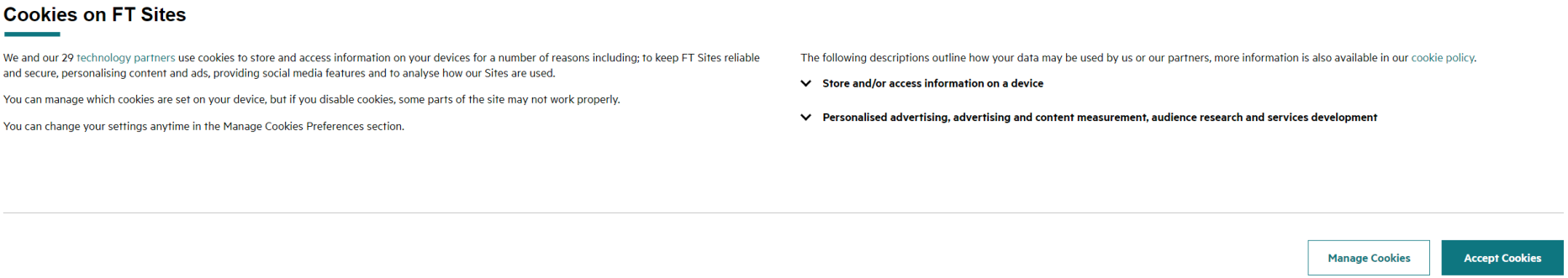
➢ Tracking users and their online habits
- Advertising companies are highly profitable
- Privacy-intrusive to the users

➢ Forms of Tracking
- Client-side
  - Cookies
- Server-side
  - User accounts
  - Device fingerprinting

➢ What to track?
- Daily expenses
- Life pattern
- Preference on commodities
- Location
- Political choice
- Religion
- …

Cookies

- HTTP protocol is stateless
- Cookies
  - **Allows a web server to store a small amount of data on the computers of visiting users, which is then sent back to the web server upon subsequent requests**
  - Today, a core technology on which complex, stateful web applications are built



Cookies on FT Sites

We and our 29 technology partners use cookies to store and access information on your devices for a number of reasons including; to keep FT Sites reliable and secure, personalising content and ads, providing social media features and to analyse how our Sites are used.

You can manage which cookies are set on your device, but if you disable cookies, some parts of the site may not work properly.

You can change your settings anytime in the Manage Cookies Preferences section.

The following descriptions outline how your data may be used by us or our partners, more information is also available in our cookie policy.

Store and/or access information on a device

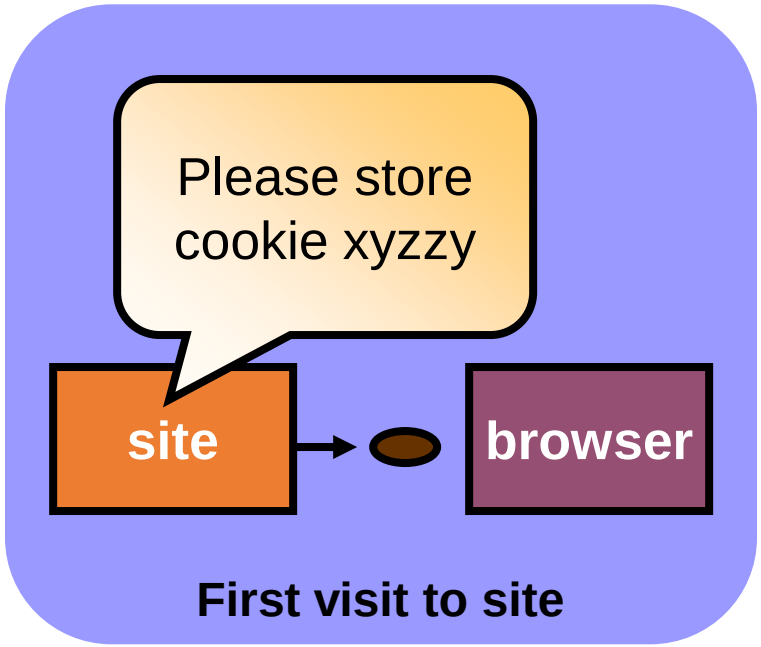
Personalised advertising, advertising and content measurement, audience research and services development

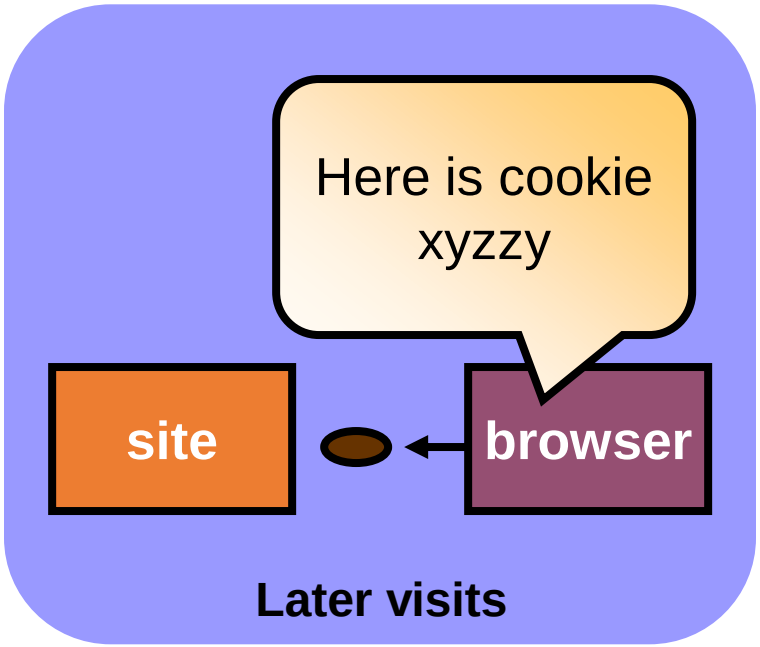
Manage Cookies

Accept Cookies

How Cookies Work – the Basics

- A cookie stores a small string of characters
- A web site asks your browser to “set” a cookie
- Whenever you return to that site your browser sends the cookie back automatically





But … Cookie Abuse?

- Can be used to track users

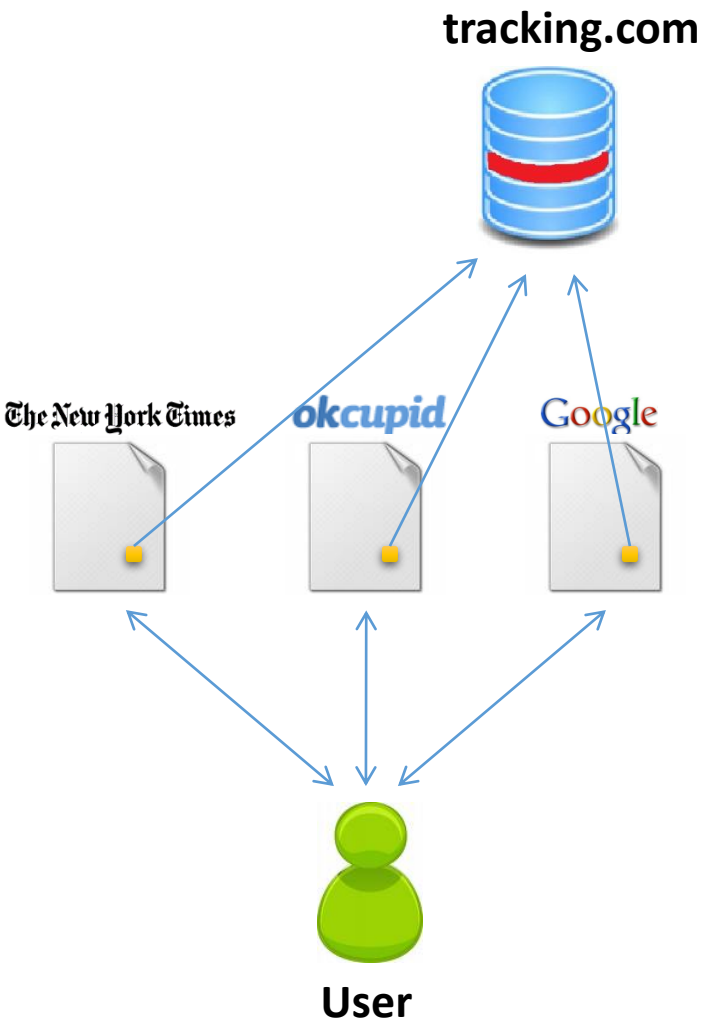
- A webpage contains various resources
  - HTML, images, CSS, JavaScript
  - Located on the hosting server, or a third-party server
    - Third-party cookies

Cookie Terminology

- **Cookie Replay** – sending a cookie back to a site
- **Session cookie** – cookie replayed only during current browsing session
- **Persistent cookie** – cookie replayed until expiration date
- **First-party cookie** – cookie associated with the site the user requested
- **Third-party cookie** – cookie associated with an image, ad, frame, or other content from a site with a different domain name that is embedded in the site the user requested
  - Browser interprets third-party cookie based on domain name, even if different domains are owned by the same company

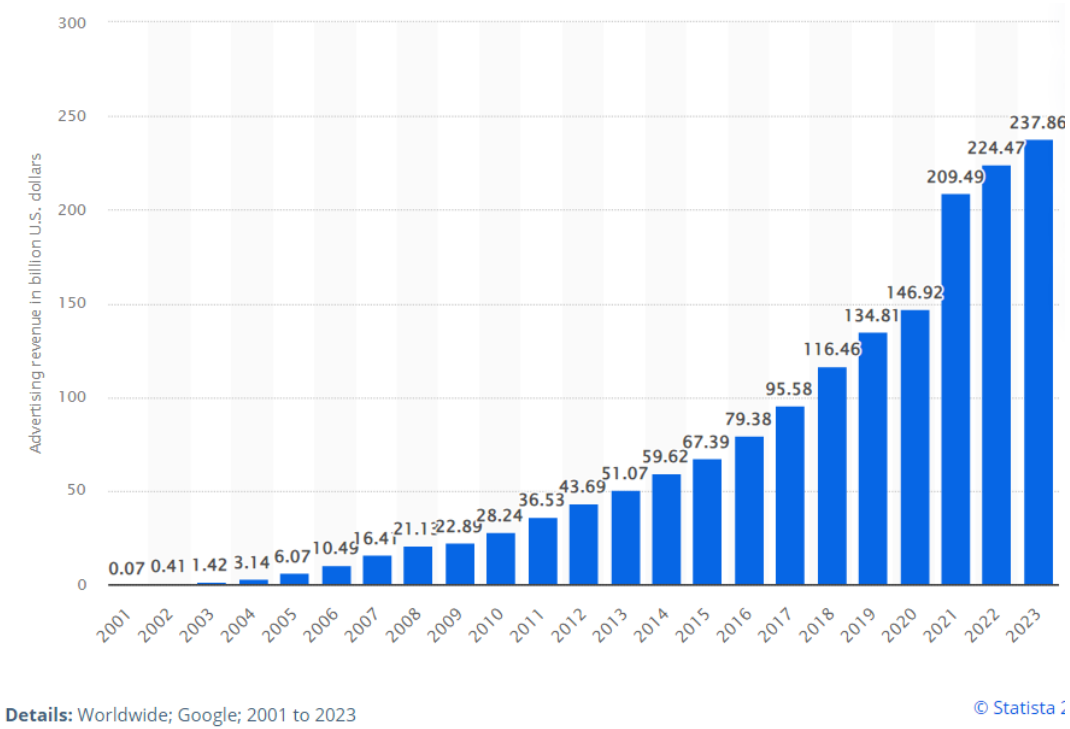
Third-party Cookies

- NYTimes.com contains an image from tracking.com
  - When downloaded, creates a cookie
- When user goes to okcupid.com, also containing a tracking.com image, tracking.com detects the user
- Allows tracking.com to profile the user
  - Mostly done by Ad networks for **behavioral advertising**



Advertising

- Online Advertising plays a critically important role in the Internet world.
- Advertising is the main way of profiting from the Internet, the history of Internet advertising developed alongside the growth of the medium itself



https://www.statista.com/statistics/266249/advertising-revenue-of-google/

Parties

- **Advertiser**
  - Got money, wants publicity
  - e.g., Coca-Cola
- **Publisher**
  - Got content, wants money
  - Cnn.com
- **Ad-network**
  - Got advertising infrastructure, wants money
  - e.g., Google AdSense
- **Consumer**
  - Wants free content

Business Model

https://www.webfx.com/how-much-does-social-media-advertising-cost.html, as in 2023

- CPM = Cost Per thousand impressions
  - Impression: user just sees the ad.
- CPC = Cost Per Click
  - This is the cost charged to an advertiser every time their ad is "clicked" on

How Ad-Networks Match Ads

- Most behavioral targeting systems work by categorizing users into one or more audience segments.
- Profiling users based on collected data
  - Search history – analyzing search keywords
  - Browse history - analyzing content of visited pages
  - Purchase history
  - Social networks
  - Geography

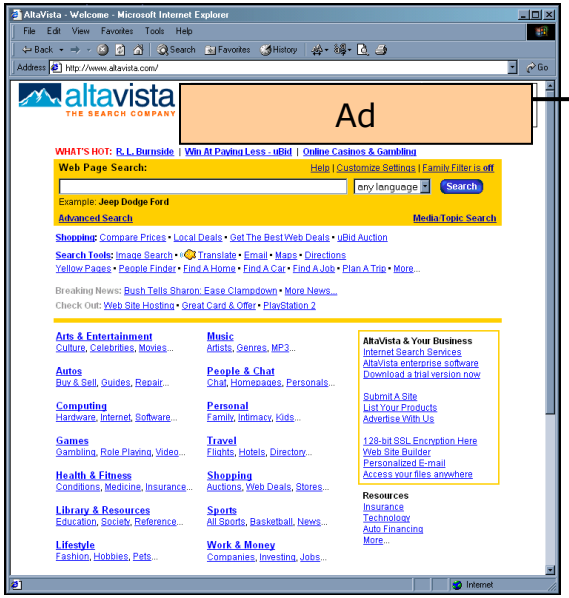
Ad Networks

search for medical information

buy CD

set cookie

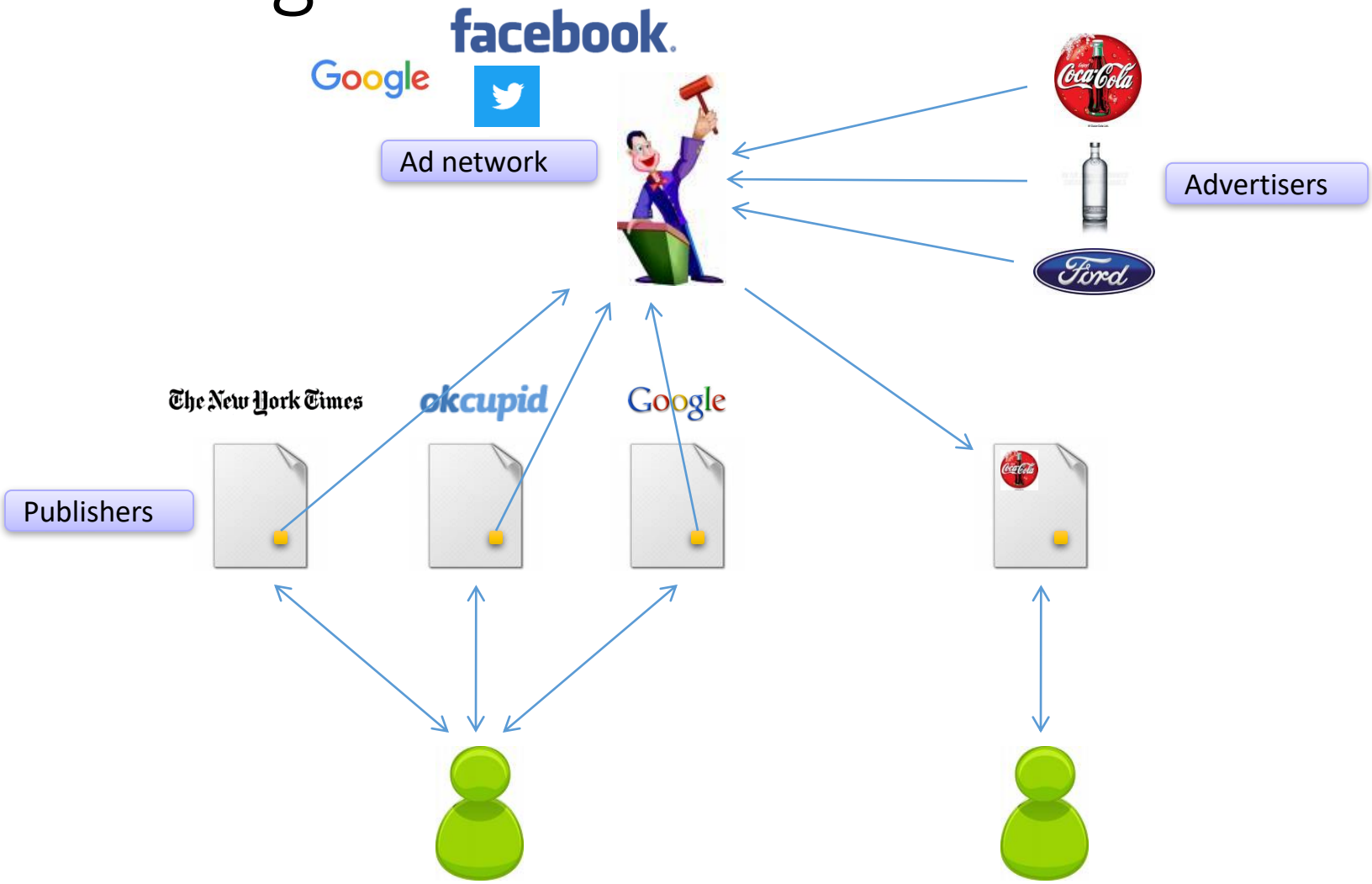
replay cookie



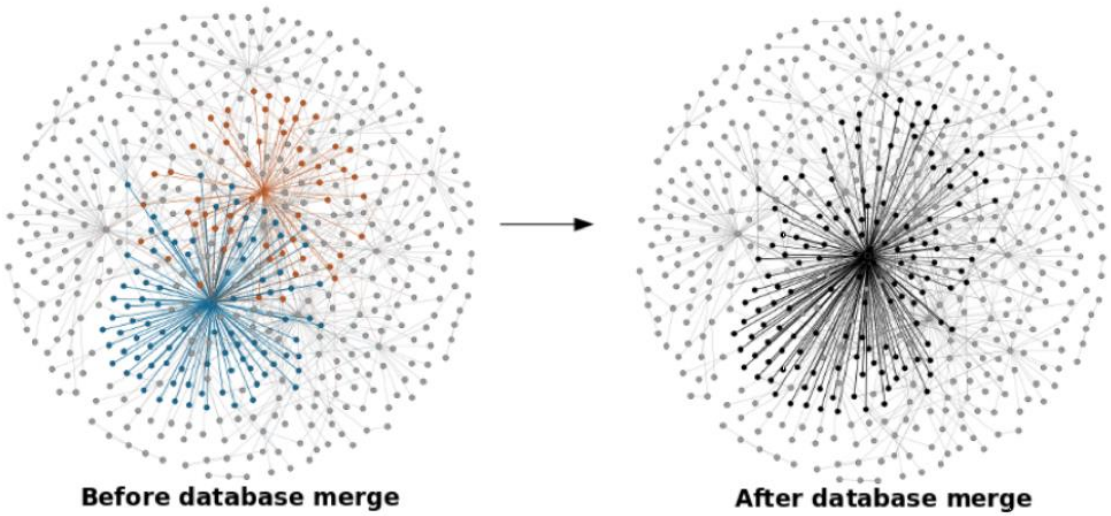
Search Service

CD Store

Behavioral Targeting: A Walk-through



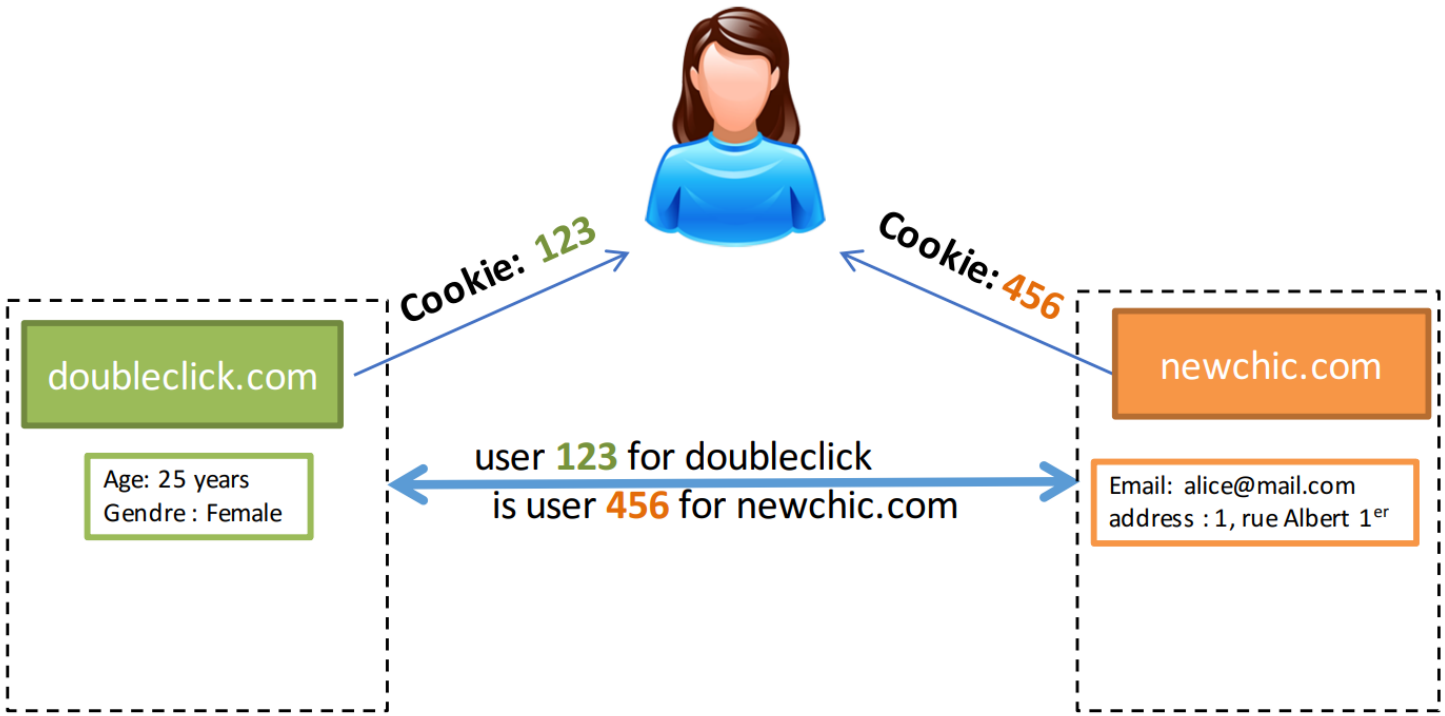
Cookie Sync: Merging Audience Data



The process by which two different trackers link the IDs they've given to the same user

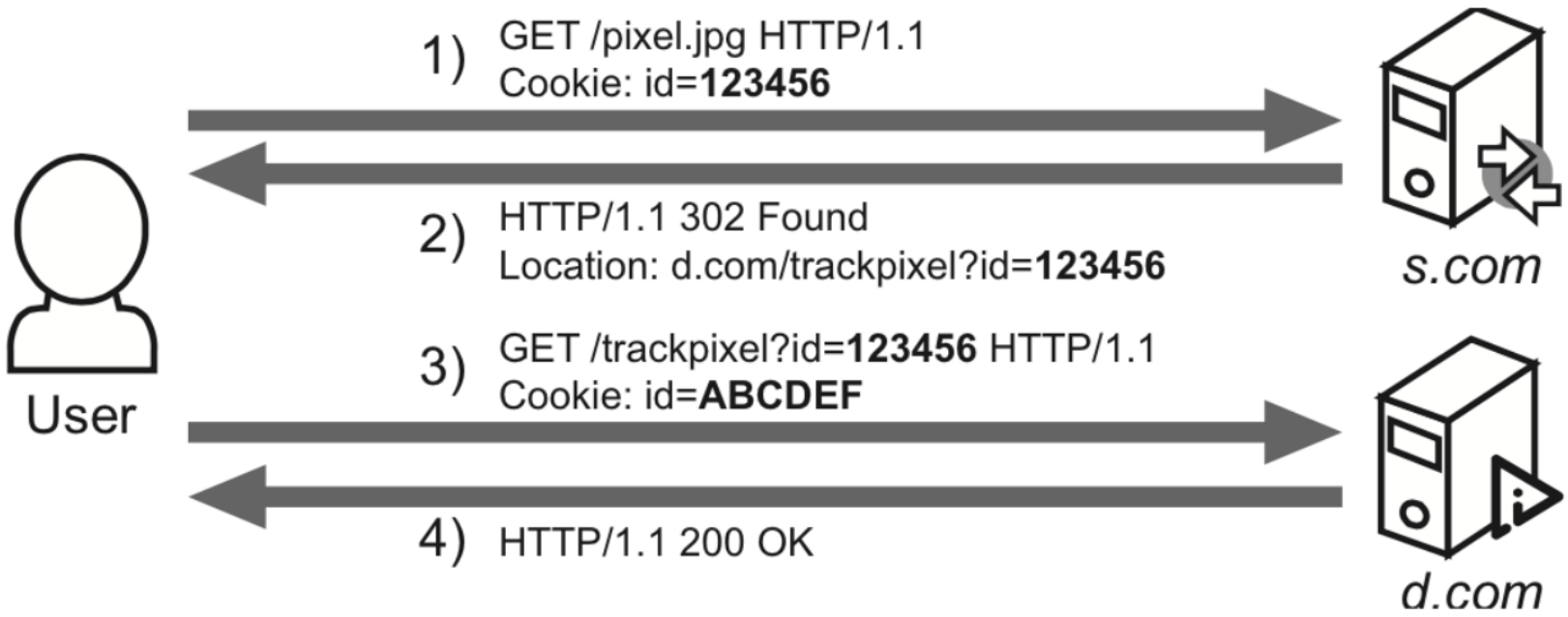
https://www.youtube.com/watch?v=-nt_fkdAUGg

Cookie Sync


Cookie: 123
Cookie: 456
doubleclick.com
newchic.com
Age: 25 years
Gendre : Female
user 123 for doubleclick
is user 456 for newchic.com
Email: alice@mail.com
address : 1, rue Albert 1er

Cookie Sync (cont.)

s.com matches their cookie with d.com using an HTTP redirect



https://www.youtube.com/watch?v=-nt_fkdAUGg

What Ad Networks may Know…

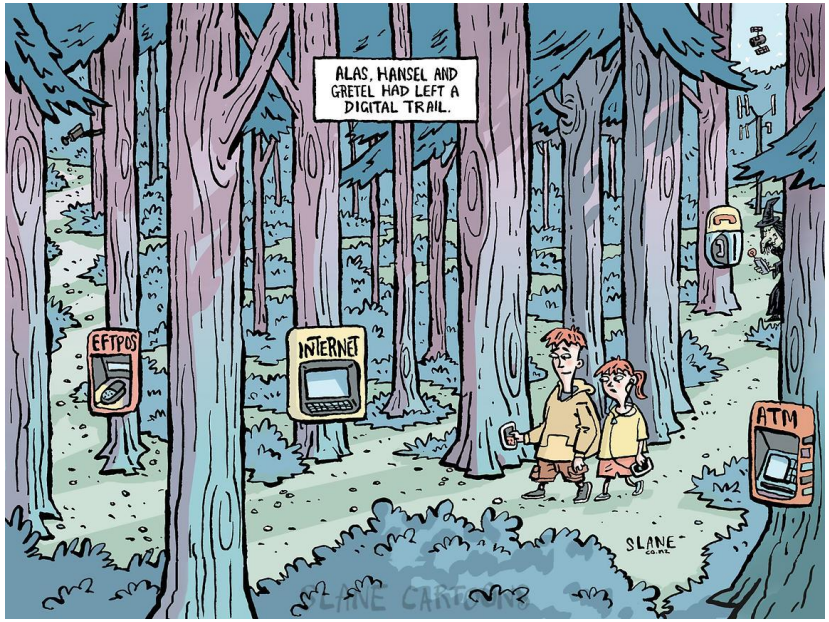
- Personal data:
  - Email address
  - Full name
  - Mailing address (street, city, state, and Zip code)
  - Phone number
- Transactional data:
  - Details of plane trips
  - Search phrases used at search engines
  - Health conditions

"Surveillance is the Business Model of the Internet"

– Bruce Schneier

Countermeasures?

- Deleting cookies frequently
  - One out of three users do it every month
- Browser extensions that reveal third-party tracking
- Modern browsers have native support to reject all third-party cookies
- Private browsing mode



Courtesy: Chris Slane

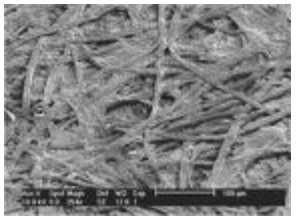
Counter-countermeasure (!) By Trackers

- Counter-countermeasure: cookie-less tracking
  - Browser (device) fingerprinting

Everything Has a Fingerprint







Browser Fingerprinting

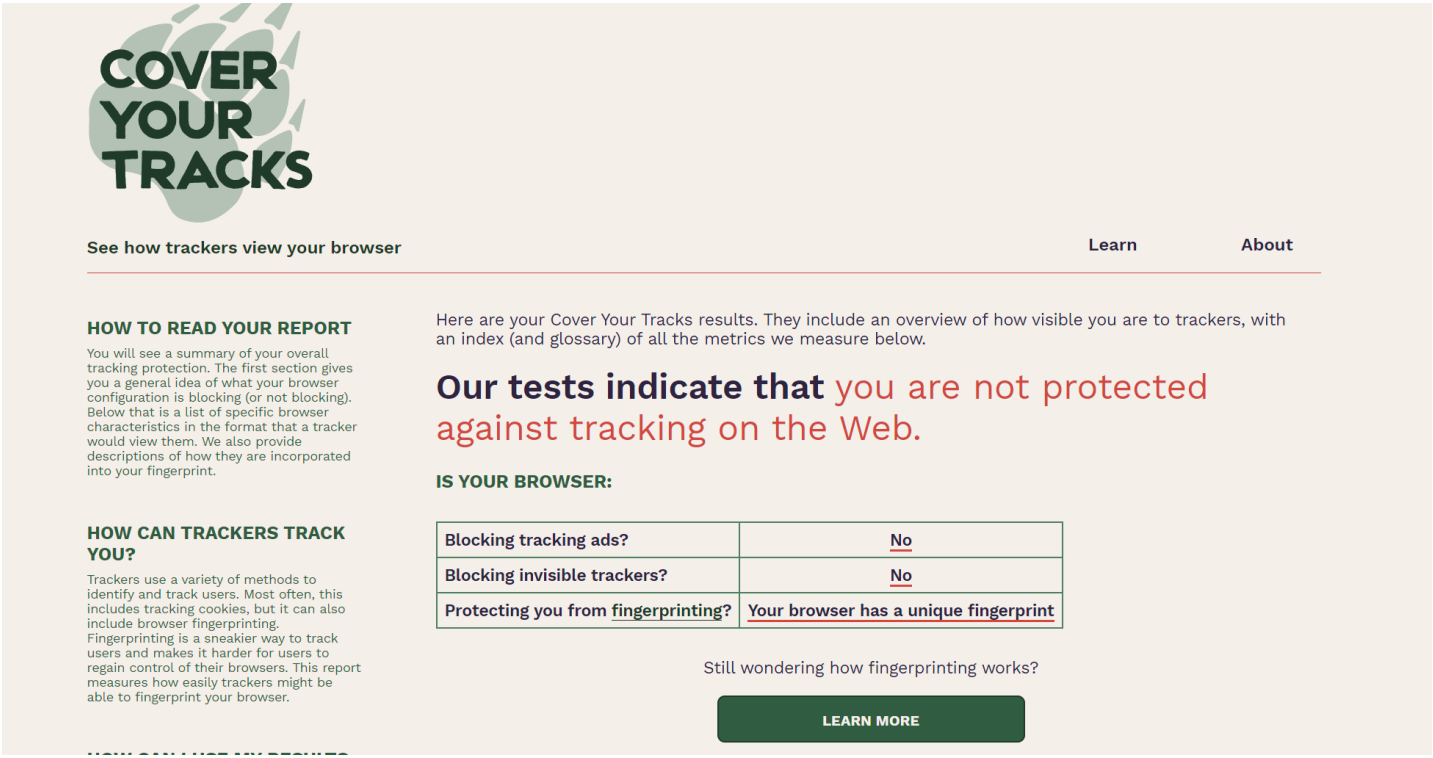
- Features of a browser and its plugins
- Introduced in 2009
- Small number of commercial companies use such methods to provide “device identification” through web-based fingerprinting
- Constructive use:
  - Combat fraud
- Destructive use:
  - Track users between sites, without their knowledge and without a simple way of opting-out
  - Deliver tailored malware

Fingerprinting Web Browsers

- User agent
- HTTP ACCEPT headers
- Browser plug-ins
- Clock skew
- Installed fonts
- Cookies enabled?
- Browser add-ons
- Screen resolution

Try your own browser

- First demonstrated in P. Eckersley, “How Unique Is Your Browser?
- Try your own browser here



Anonymity Networks

Privacy on Public Network

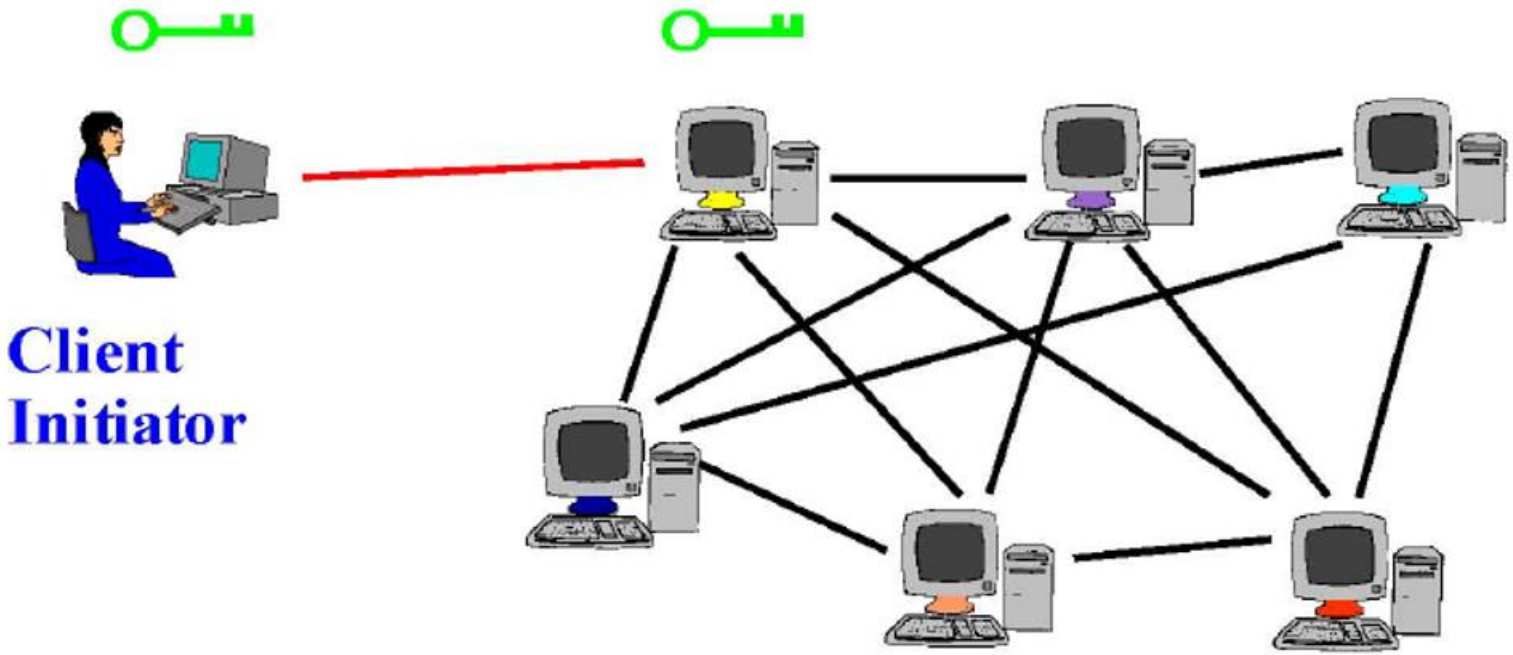
- Internet is designed as a public network
- Routing information is public
  - IP packet header identify source and destination.
  - Even a passive observer can easily figure out who is talking to whom.
- Encryption does not hide identities
  - Encryption hides payload, but not routing headers.
  - Even IP-level encryption (e.g., VPN) reveal IP address of gateways.

The Onion Router (Tor) Network

- https://tor.eff.org/
- Tor is designed to hide users' identities and their online activity from surveillance and traffic analysis by separating identification and routing.
- Initially, TOR aimed to provide a secure and anonymous communication method for the U.S. government.
- Now, TOR evolves into an open-source anonymous network project.
- Main idea: Passing the data through a circuit of at least three different routers. Each router only knows its **predecessor** and **successor**.

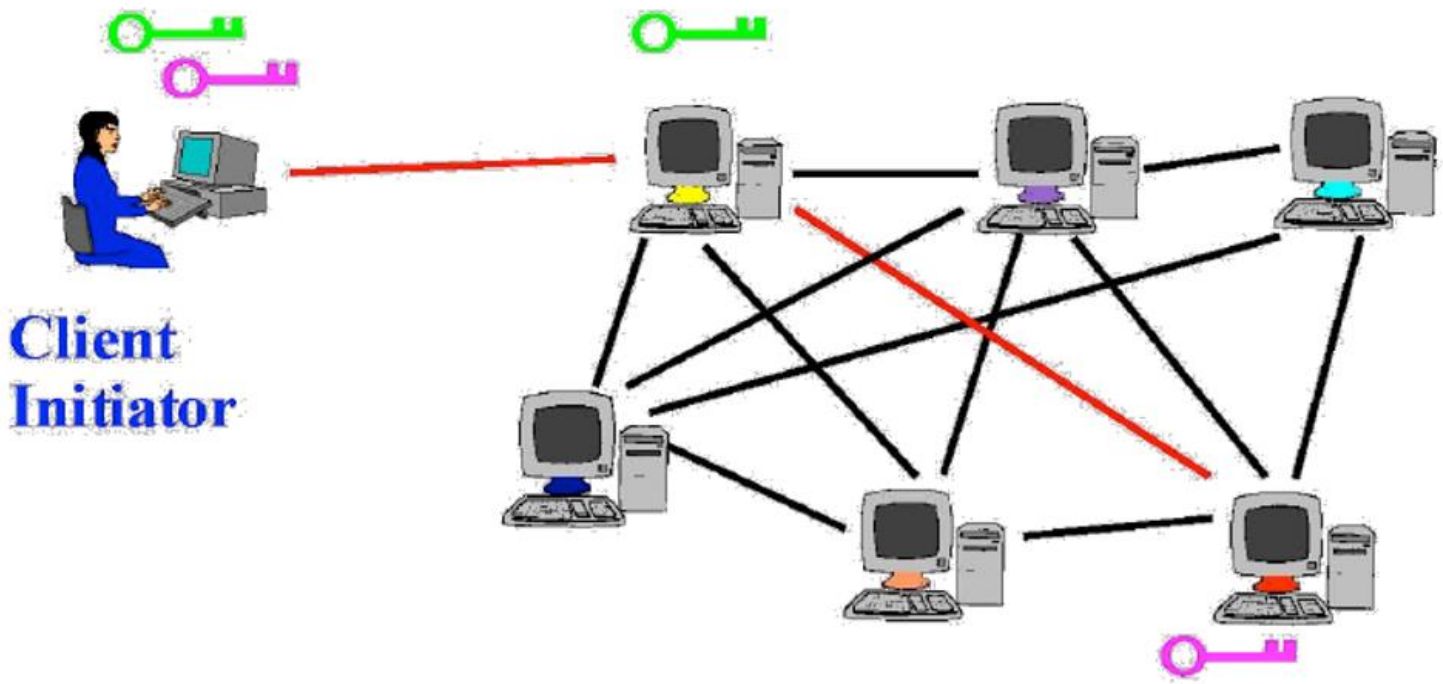
Tor Circuit Setup (1)

- The Tor client selects an Onion router as the entry node (Onion Router #1) and establishes a symmetric session key with the selected entry node (circuit with entry node).



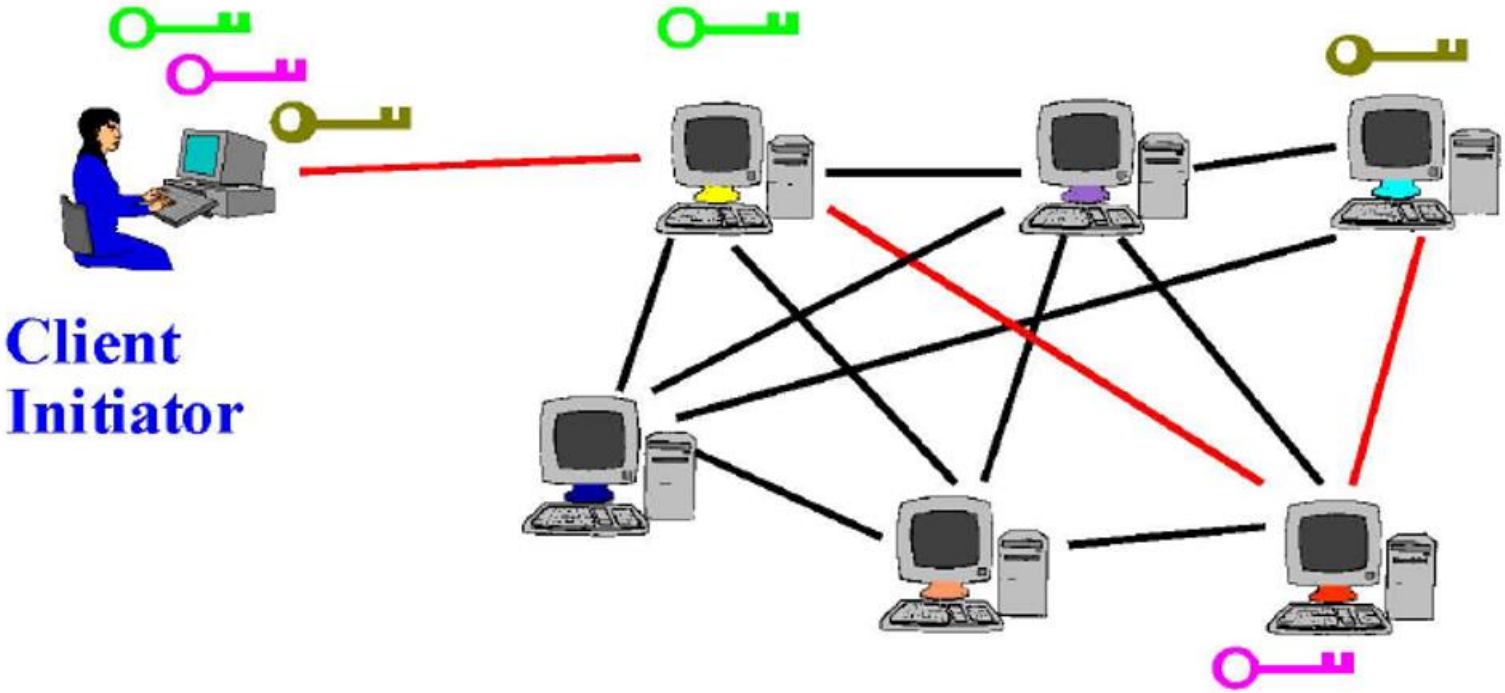
Tor Circuit Setup (2)

- The Tor client extents the circuit by establishing a symmetric session key with Onion Router #2
- Tunnel through Onion Router #1



Tor Circuit Setup (3)

- The Tor client extents the circuit by establishing a symmetric session key with Onion Router #3 (also known as exit node).
- Tunnel through Onion Router #1 and #2



Using a Tor Circuit

- The client connects and communicates over the established Tor circuit.
- The client encrypt the datagram using the shared session keys (3 times in this example). The datagram ciphertext is decrypted after passing through each router in the circuit.