CookieGraph and FP-Fed

Advanced Computer Security
CS563/ECE524

Nikita Borisov, Spring 2025

Today

- CookieGraph: 1st party tracking cookies
- FP-Fed: detecting fingerprinting scripts

Logistics

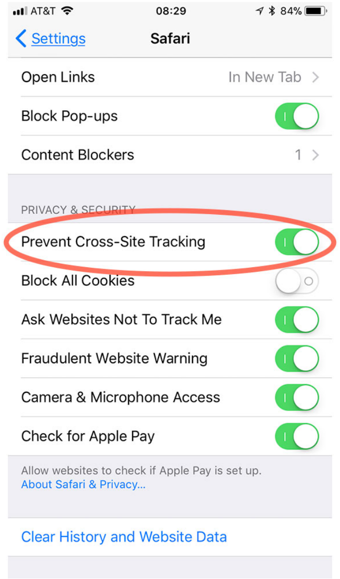
- Schedule updated
  - Slight changes
  - Network security
- Volunteer to blog, present
  - FCFS!
  - Blog due 2 weeks after lecture
- Don’t forget to submit reviews

Cookie Graph

- Measurement of 1st party cookies that are used when 3rd party cookies are blocked
- ML-based countermeasure to block *tracking* while leaving *functional* cookies
- Fingerprinting scripts are used to set 1st party cookies

3rd Party Cookies

- 3rd party cookies are used to profile users *across sites*
- Many browsers + extensions block 3rd party cookies
- Safari: Block cookies on 3rd party sites w/o 1st party relationship
- Safari ATP,. Edge: block cookies based on a list
- Firefox ETP: partition state based on 1st party
- Chrome: no blocking, plan to give users *option* to block cookies



Safari

Open Links — In New Tab

Block Pop-ups ●

Content Blockers — 1

PRIVACY & SECURITY

Prevent Cross-Site Tracking ●

Block All Cookies ○

Ask Websites Not To Track Me ●

Fraudulent Website Warning ●

Camera & Microphone Access ●

Check for Apple Pay ●

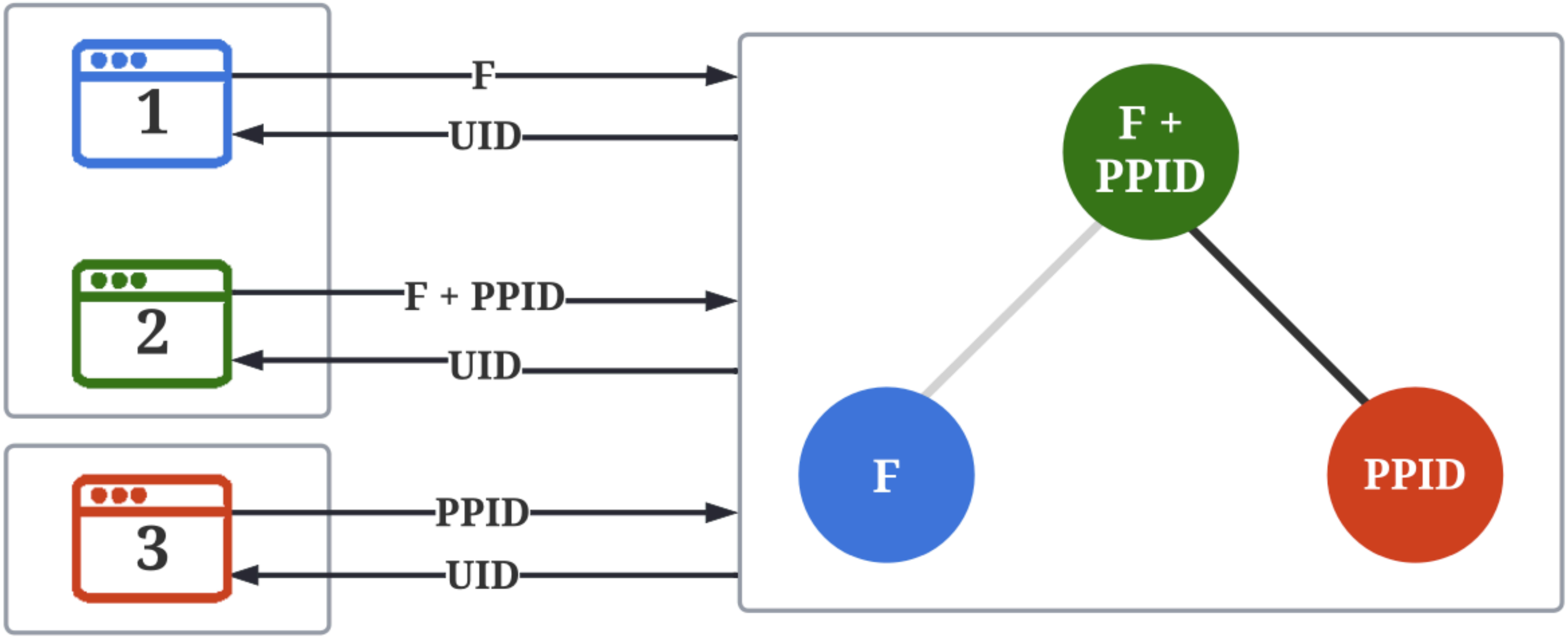
Allow websites to check if Apple Pay is set up.
About Safari & Privacy…

Clear History and Website Data

1st party cookies

- 1st party can be used for cross-site tracking
  - Link cookies to PII (email, etc.)
  - Cookie Sync: share identifier in request
  - Fingerprinting: set cookie based on FP
  - CNAME cloaking: create a 1st-party URL for 3rd party
- Countermeasures:
  - Safari has a blocklist

Identifier Graph



Differential Study

- Crawl with / without 3rd party blocked
- Measure # of requests to *known* trackers
  - With & without identifier included in request
- Conclusions
  - 3rd party blocking has minimal effect on requests to trackers
  - Ergo 1st party cookies must be a major contributor (?)
- Thoughts?

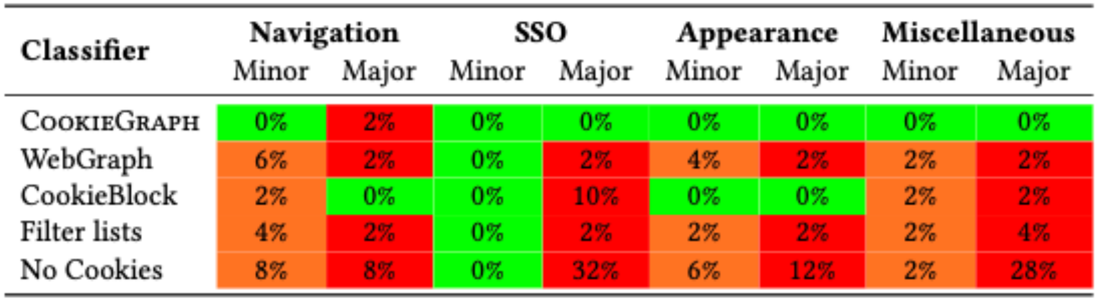
CookieGraph

- Graph of cookie data flow
  - Set by servers (infiltration)
  - Stored in cookies
  - Used by scripts
  - Sent to trackers (exfiltration)
- Extract features
  - Structural, flow

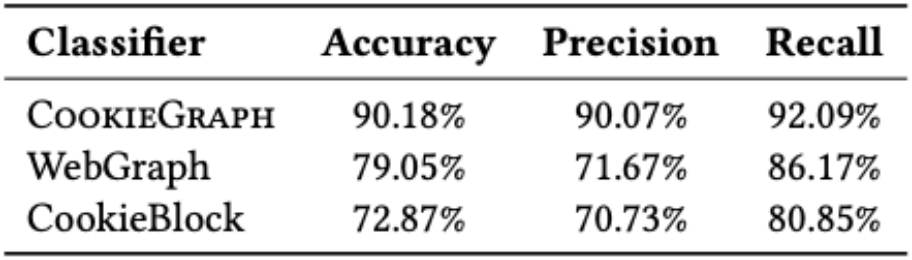
Training

- Train using ground truth data sets
  - Cookiepedia
- Use features to classify cookies as functional / tracking
  - Flow features more important

Results



Classifier	Navigation		SSO		Appearance		Miscellaneous	
	Minor	Major	Minor	Major	Minor	Major	Minor	Major
CookieGraph	0%	2%	0%	0%	0%	0%	0%	0%
WebGraph	6%	2%	0%	2%	4%	2%	2%	2%
CookieBlock	2%	0%	0%	10%	0%	0%	2%	2%
Filter lists	4%	2%	0%	2%	2%	2%	2%	4%
No Cookies	8%	8%	0%	32%	6%	12%	2%	28%



Classifier	Accuracy	Precision	Recall
CookieGraph	90.18%	90.07%	92.09%
WebGraph	79.05%	71.67%	86.17%
CookieBlock	72.87%	70.73%	80.85%

Limitations

- Offline detection and labeling for list generation
- Single browser platform
- Ground truth
- Breakage classification
- Performance

Evaluation

Positive Points

- **High accuracy and minimal site breakage:** CookieGraph detects first-party tracking cookies with 90.18% accuracy while maintaining functionality.
- **Innovative and robust approach:** Uses graph-based machine learning focused on tracking behaviors, making it resilient to manipulation.
- **Comprehensive evaluation:** Large-scale study highlights the widespread use of first-party cookies and demonstrates CookieGraph's superiority over prior methods.

Areas for Improvement

- **Performance optimization:** Reduce overhead to enable real-time deployment and improve efficiency for broader adoption.
- **Broader testing and data expansion:** Test across multiple browsers/platforms and enhance ground truth labels with crowdsourced or regulatory data.
- **Defense against adversarial attacks:** Strengthen the model to handle evolving tracking methods, including graph manipulation and alternative storage mechanisms.

Discussion Points

- Viability of offline blocklists vs online detection
- How does this interact with:
  - Cookie consent dialogs?
  - Privacy regulation?
- How will tracking evolve if CG is widely adopted
  - Interfere with data collection
  - Adversarial ML
  - Other tracking techniques?

More discussion / takeaways?

- Anyone really like / really dislike the paper? Why or why not?
- Anything surprising?