How Broadcast Data Reveals Your Identity and Social Graph

Rolf Winter <rolf.winter@hs-augsburg.de>
Michael Faath <michael.faath@hs-augsburg.de>
Fabian Weisshaar <fabian.weisshaar@hs-augsburg.de>

Idea

- Connect to a large network and analyse everything received
  - Excluding the traffic the listener introduces
- Are there protocols “polluting” the network?
- What can we learn from this data?
  - Protocols
  - Devices
  - Users and groups of users

Experiment locations

- The lab
  - Controlled environment
- A wireless campus network: Eduroam
  - Over 6,000 students and staff
- IETF Meeting network
  - IETF 93 - Prague / IETF 94 - Yokohama

Legal aspects - I am not a lawyer

- IETF Meeting experiment announcement[1]
  - First reaction: “doesn't this fall under human subjects rules for experiments [...]?”[1]
  - Over 40 mailing list responses
  - Experiment might break EU data protection laws
  - But: more positive than negative reactions
- Legal questions could not be resolved in time
  - Experiment for the 93rd IETF meeting cancelled
  - → Proposal to establish the IETF Experiment Ethics Review Board[2]

[1] “Multicast/Broadcast Experiment at IETF94 (email thread),” Nov. 2015. [Online]. Available: https://www.ietf.org/mail-archive/web/94attendees/current/msg00490.html
[2] https://www.ietf.org/blog/2015/09/experiment-ethics-and-privacy/

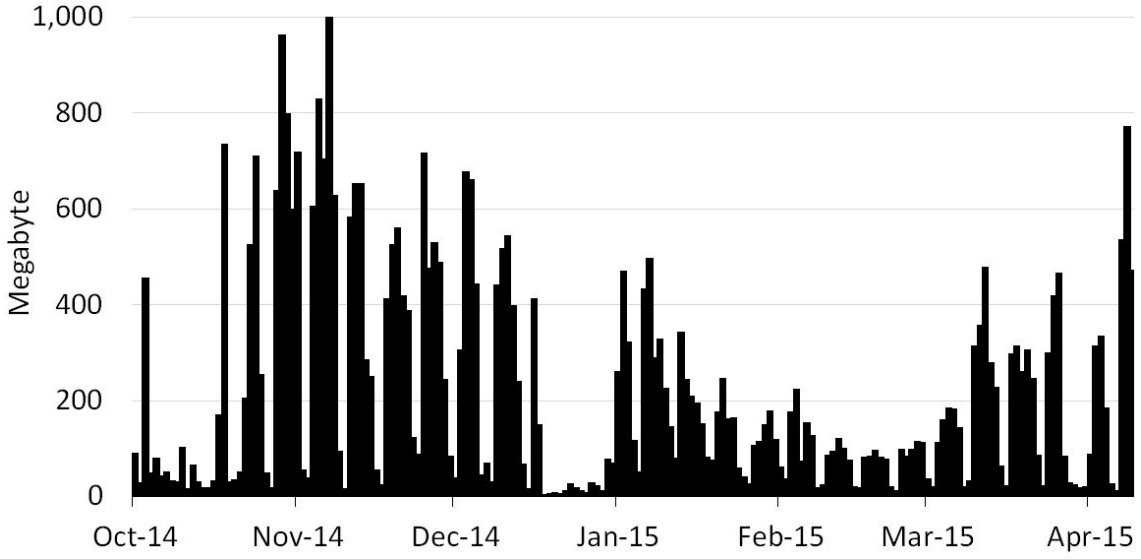
Legal aspects - I am not a lawyer

- Legal statement by the German National Research and Education Network (DFN)[1]
  - It is not okay (for universities in Germany) to store and analyze broadcast data
  - Consent of every user in the network is necessary
  - It *might* be okay to store and analyze for specific research if privacy of users is ensured
- Remove all personally identifiable information
  - MACs, IPs, hostnames etc. hashed
  - Analyzation only for selected protocols possible
  - Don’t store raw data

[1] H. Sporleder, “Dein Name ist Programm”, DFN Infobrief Recht, pp. 16–18, Nov. 2015

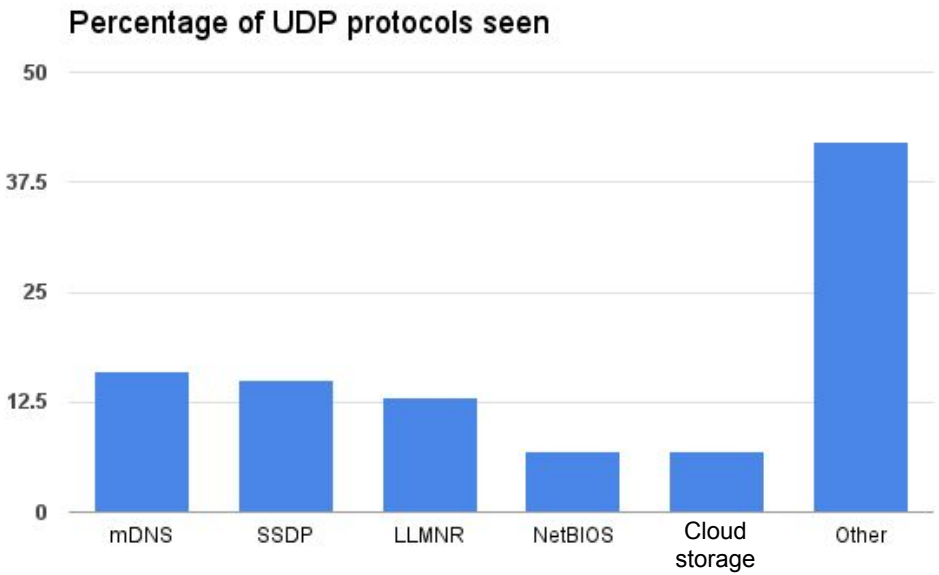
Data analysis: Campus network

- All eduroam users on campus are in one broadcast domain
  - Plus all of the VPN users from home
- Six months
- ~40 GB of data seen
  - ~215 MB per day on average



Data analysis: Campus network

- ~35,000 MAC addresses seen
  - max. 21,000 from real devices
- ~90% UDP packets
  - Focus on most seen protocols
  - Analysis of payload



Desktop app of a popular cloud storage service

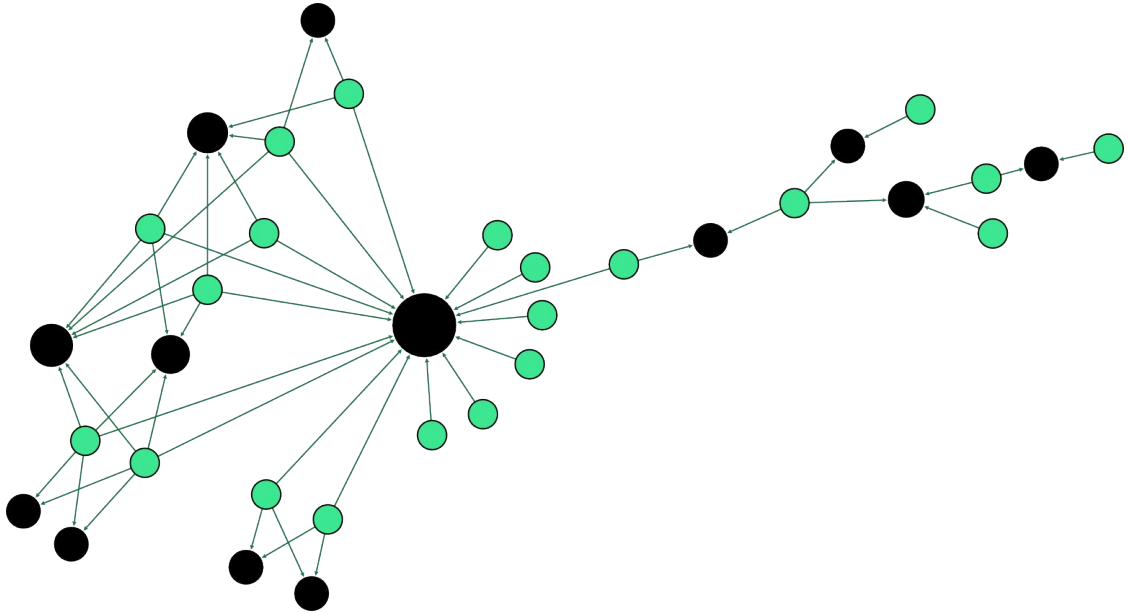
- Used to store and share data in the cloud
- Implements a protocol for local data exchange
- Broadcasts multiple packets every 30 seconds
  - *host_int*
    - Unique ID for application installation
    - Tracking of a user even if IP or MAC address changes
  - *namespaces*
    - List of unique IDs for all known shares

Data analysis: Cloud storage service

- 2,560 application installations
- 9,361 unique shares
- Students might use the application to share data from lectures
  - ...can we draw a graph from this?

Data analysis: Cloud storage - a community graph

- Identify communites (Louvain method[1])

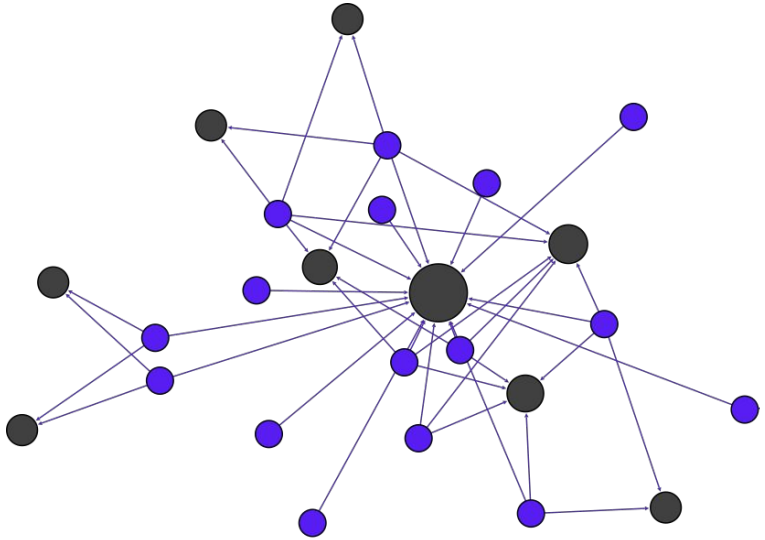


[1] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. J. Stat. Mech, 2008

Data analysis: Hostnames

- Some protocols broadcast hostnames
  - mDNS, NetBIOS, LLMNR, ...
- 7,600 hostnames found
  - removed duplicates and typical strings (“iphone”, “macbook”, ...)
  - 5,300 host names remaining
- Lots of users reveal
  - Language (“iPhone *von* John Doe”)
  - Device vendor / model (“MacBook Pro”)
  - Locations and functions (“printer”, “cs-faculty”)
  - Names (login names, nicknames, initials)

Data analysis: Hostnames



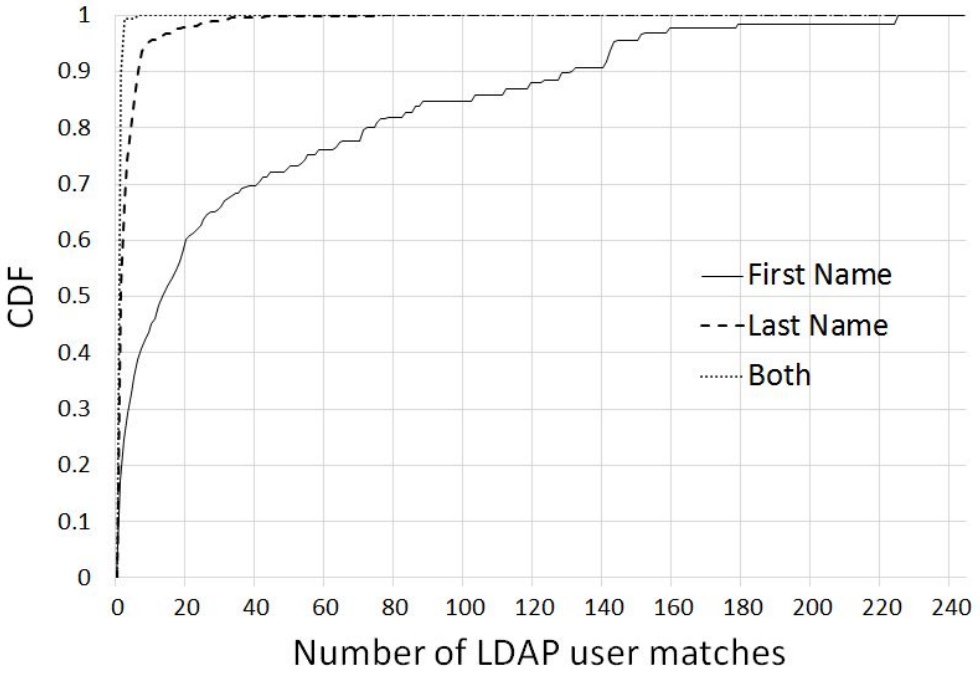
- Helps to partially identify nodes
  - But we can do more
  - If there would be a database containing all students...

Data analysis: LDAP

- LDAP server of the university is accessible from within the network
- Crawl all entries: >8,400 users
  - Login, first and last name
  - Department
  - Course of study
  - Status (student, professor, staff, ...)
  - Date of last password change
- 4,564 unique last names
- 1,300 unique first names
- Compare them to the hostnames

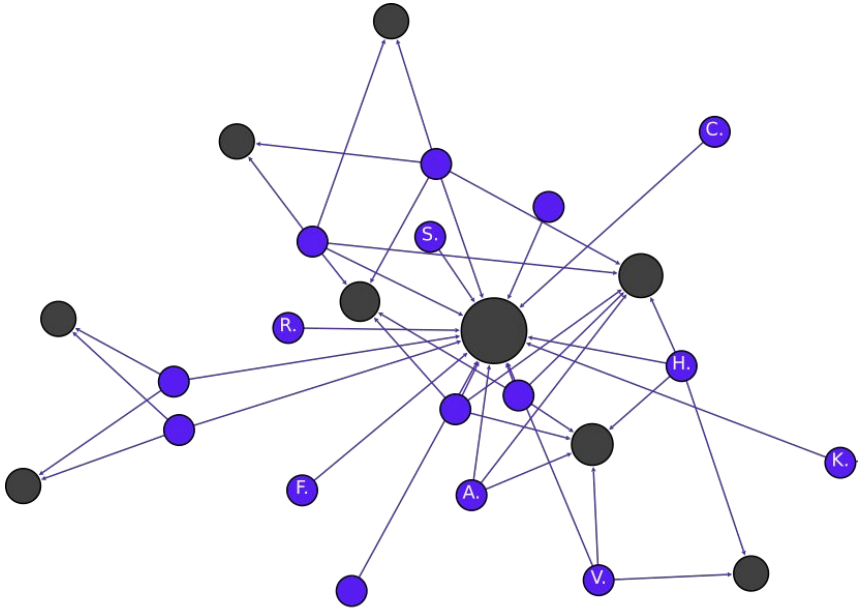
Data analysis: LDAP

- 2,900 first names matched
  - ~17% (500) match uniquely
- 929 last names matched
  - ~50% (464) match uniquely
- 293 full names matched
  - ~90% (263) match uniquely



Combining the data

- Add LDAP users to nodes
- Several users could be identified
  - Same course of studies
  - Same date for last password changed
- Those help to identify nodes with multiple LDAP matches



Data verification

- We made some surprise visits to lectures
  - Controlled experiment
  - Voluntarily data verification
- Other things to do
  - Look for social network profiles
  - Crawl the timetables of the university and match online times of the community

Countermeasures

- Don’t name your device after yourself[1]
  - Not even if it is a common nickname
- Restrict publicly visible data in your online profiles
- Switch off broadcast/multicast functionalities
  - Don’t actually do this
  - Broadcast and multicast protocols are important
- Be careful when designing broadcast protocols
  - IETF draft: Privacy considerations for IP broadcast and multicast protocol designers[2]

[1] https://tools.ietf.org/html/rfc8117
[2] https://datatracker.ietf.org/doc/draft-intarea-broadcast-consider/

Conclusion

- Personal information can be learned from broadcasts
- No protocol alone is to blame
- Check with a lawyer before doing anything like this
  - Note: criminals might not care about privacy
- Countermeasures are available and easy
  - But need a change in user behaviour