CO643 – Week 5
Social Computing

Dr Özgür Kafalı

Lecturer

R.O.Kafali@kent.ac.uk

Outline

- Social computing research area
- Online social networks
- Inference and disclosure
- Data anonymity case studies
- Legal side

University of Kent

Computing

Learning Outcomes

- After this lecture, you will be able to
  - Describe what social computing constitutes
  - Define an online social network
  - Describe the important computing problems surrounding social networks
  - Understand how you should protect the privacy of your users
  - Review legal issues for social networks

University of Kent

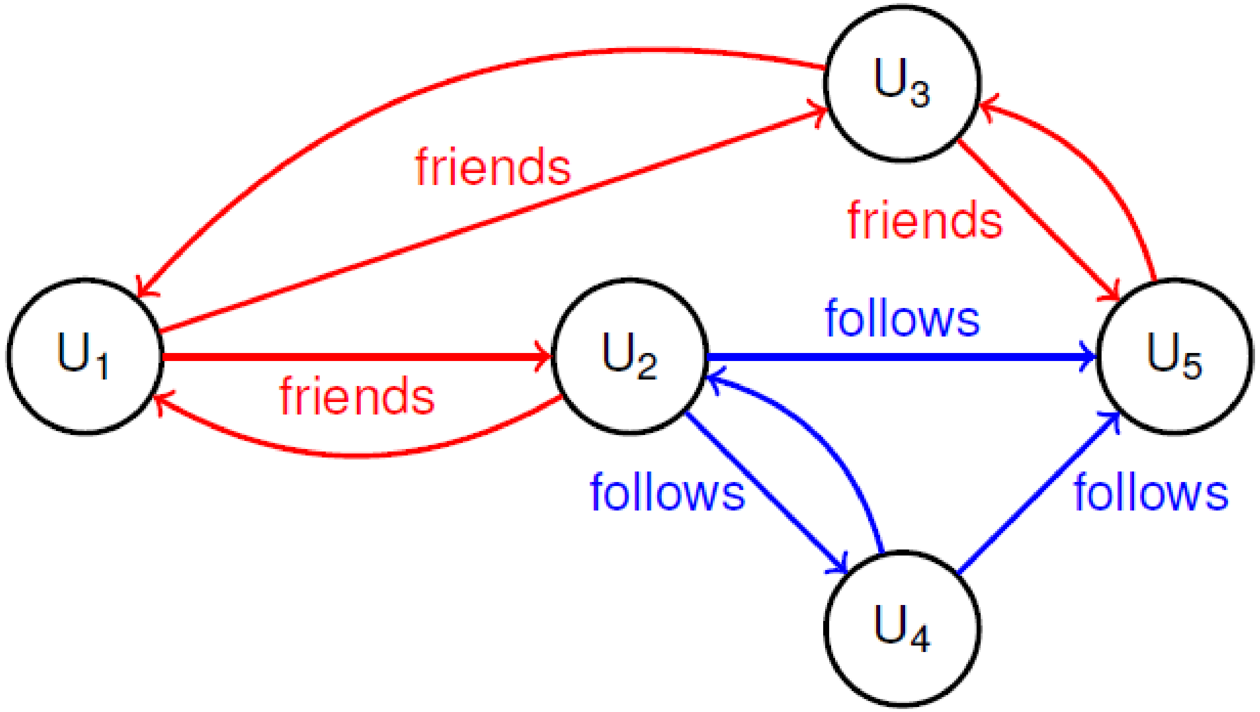
Computing

Social Computing

- An area of computer science
  - Social behavior
  - Computational methods
- Methodology
  - Models of social relationships
  - Interactions among social entities
- Application area: Online Social Networks (OSN)

OSN Basics

Directed graph G = (V, E)



Data Collection, Storage and Usage

- Collection: What personal information is collected by organisations?
- Storage: How do organisations store personal information? Is it kept secure?
- Usage: How do organisations use personal information?
  - Whom do they share it with?
  - Do they make users aware, e.g. ask for consent?

Problems

- Inference
- Sharing and disclosure
- Data anonymity
- Conflicting policies

Logic Inference

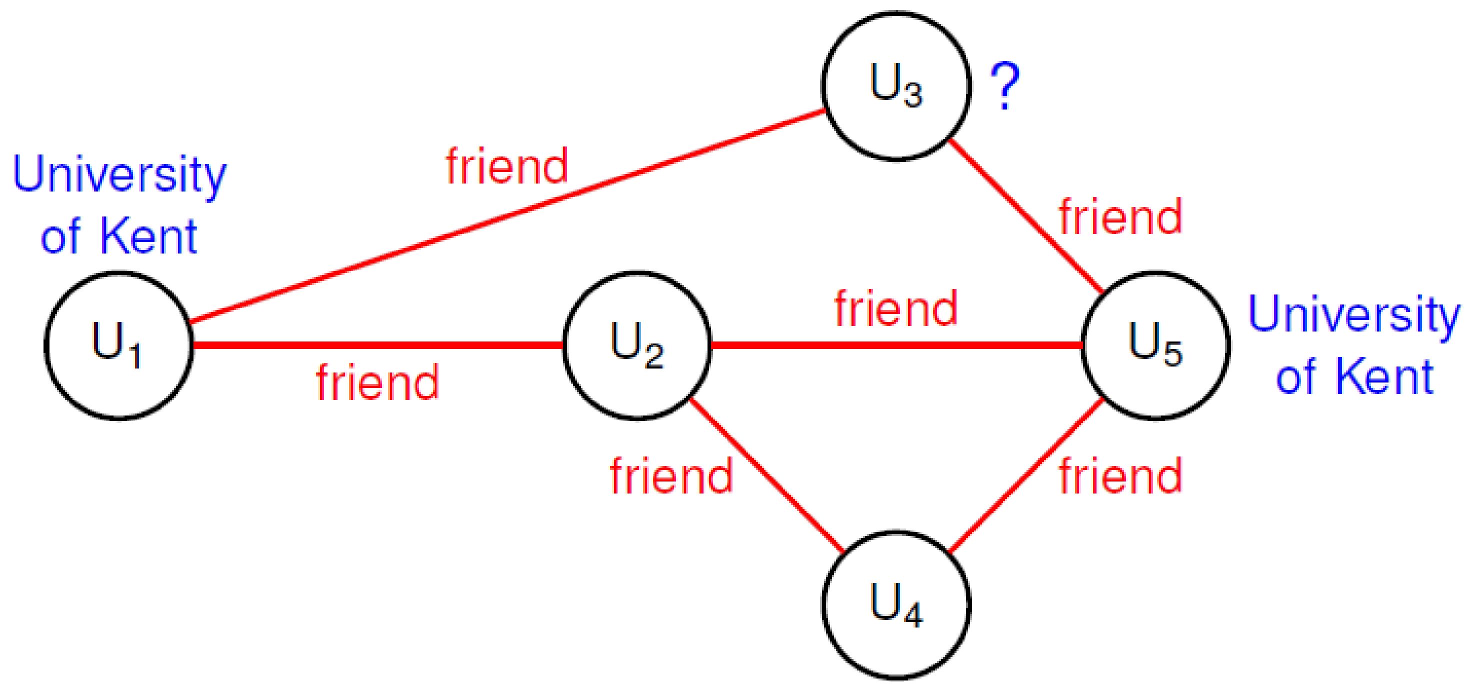
- "The act or process of deriving logical conclusions from premises known or assumed to be true"
  - Example in first order logic
    - All humans are mortal. $\forall X: human(X) \rightarrow mortal(X)$
    - All Greeks are humans. $\forall X: greek(X) \rightarrow human(X)$
    - Therefore, all Greeks are mortal. $\forall X: greek(X) \rightarrow mortal(X)$

The American Heritage Dictionary of the English Language, Fifth Edition, 2017

Inference

What can you infer about the man in the picture?

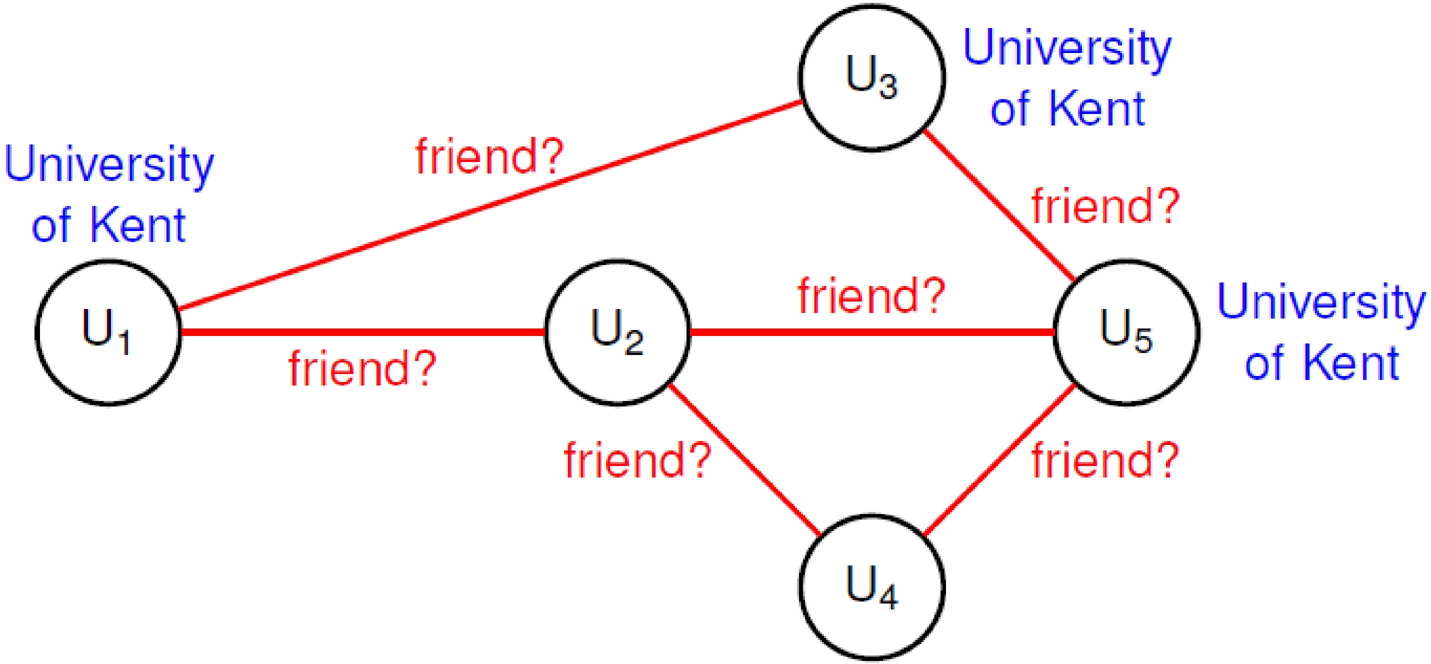
Inferring User Attributes



- Can we infer missing attributes of a user based on other users' attributes and their social links?

Mislove et al. You Are Who You Know: Inferring User Profiles in Online Social Networks Conference on Web Search and Data Mining, pages 251–260, 2010

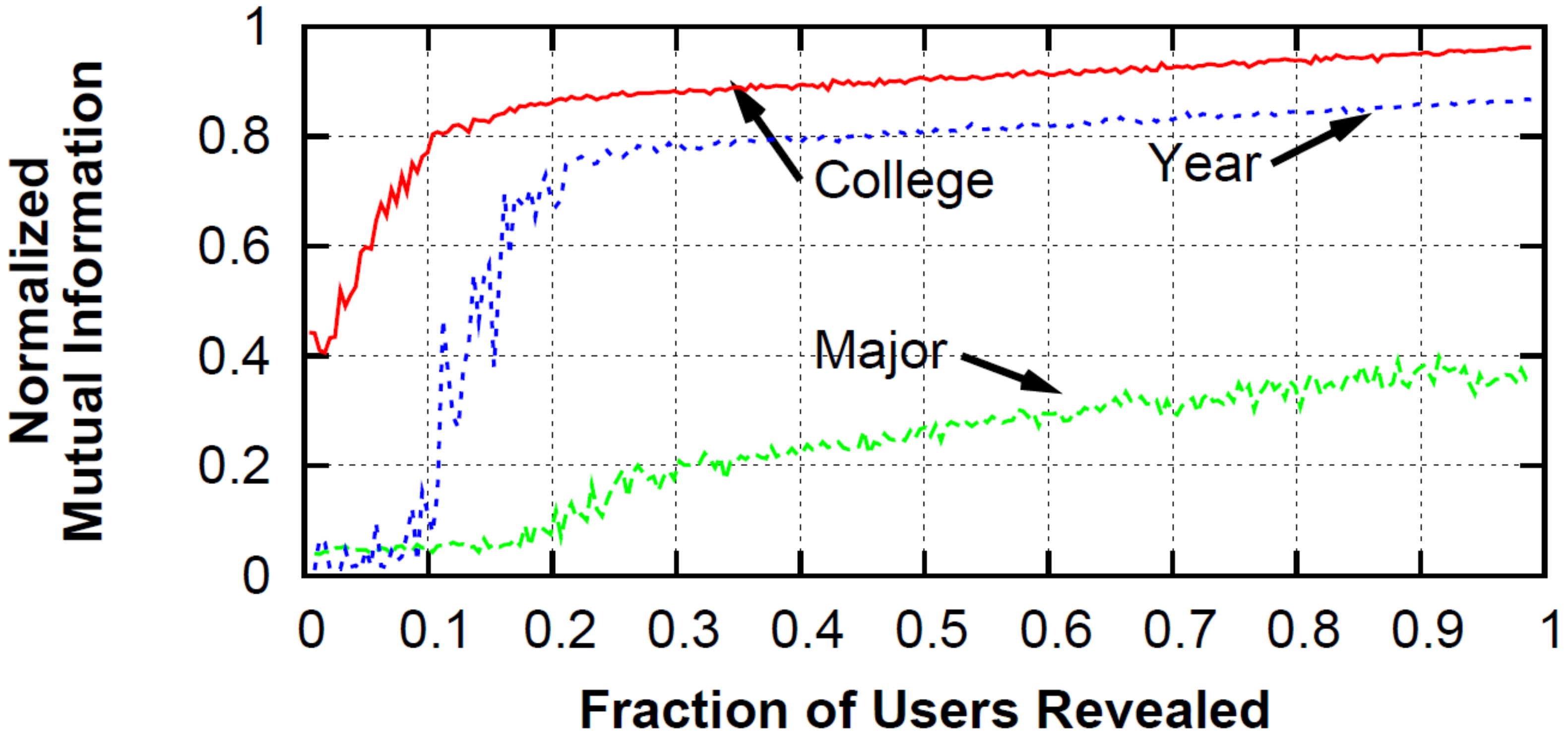
Inferring User Relations



- Can we infer social links among users based on attributes of users?

Mislove et al. You Are Who You Know: Inferring User Profiles in Online Social Networks Conference on Web Search and Data Mining, pages 251–260, 2010

Significant Outcomes



Mislove et al. You Are Who You Know: Inferring User Profiles in Online Social Networks
Conference on Web Search and Data Mining, pages 251–260, 2010

University of Kent Computing

Implications

- Probabilistic inference
  - Machine learning: Quite high accuracy with large amounts of data
  - Probabilistic model checkers: Facts and assumptions
- Better recommender systems
- Connect people who might benefit from the interaction, e.g. job search

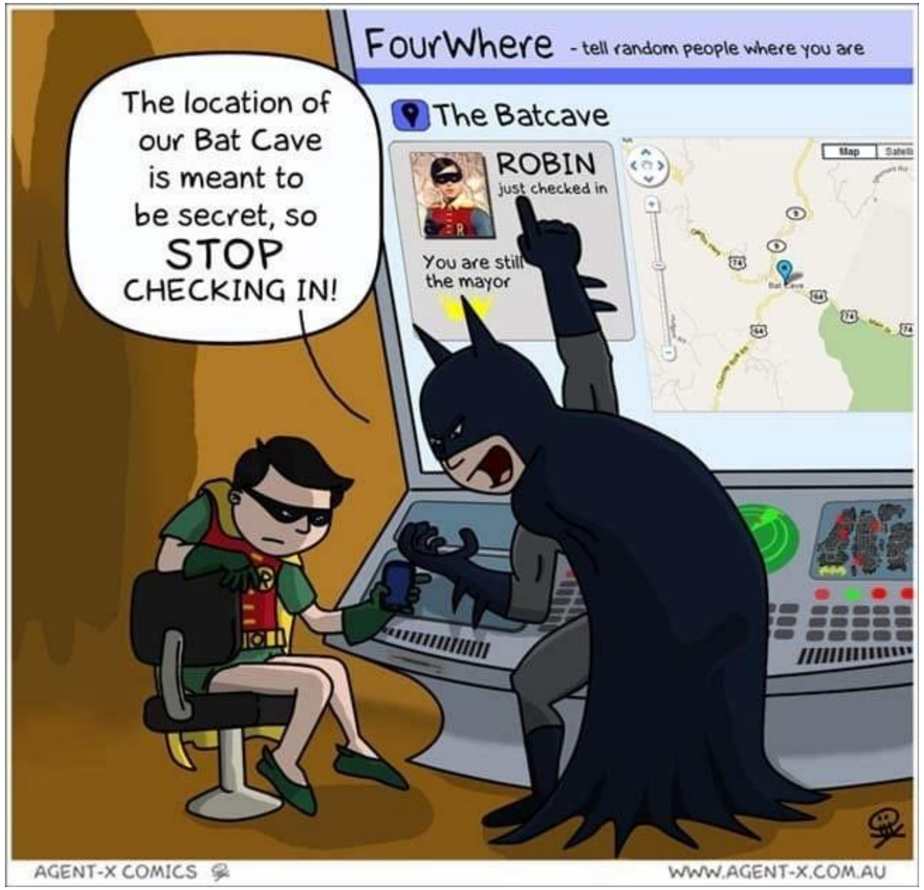
- A user's privacy no longer depends only on what they reveal
- Invasiveness of emerging machine learning technologies on user privacy

Content Sharing

What is wrong with this picture?

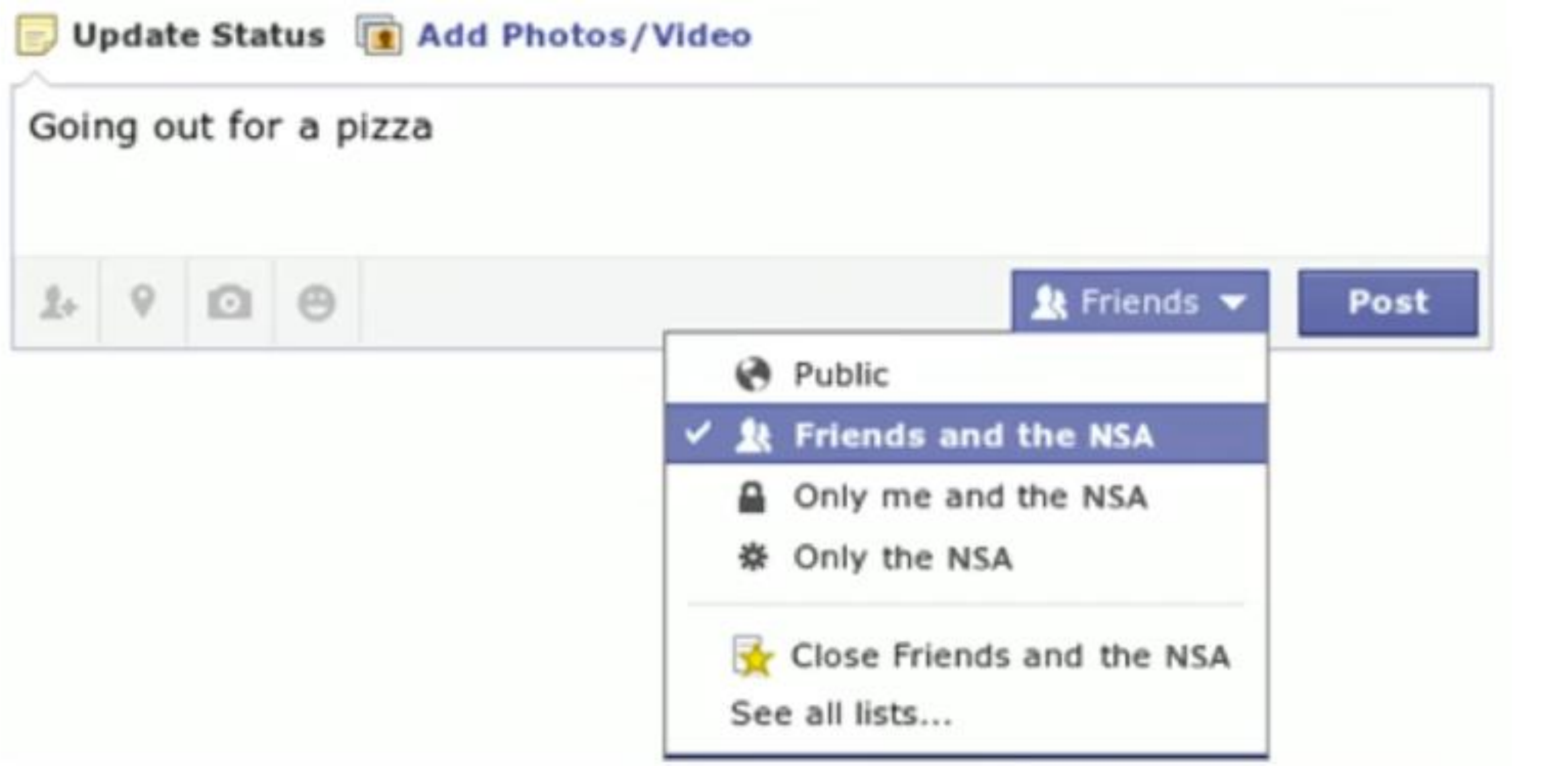
Location Sharing

- People do not really think about the consequences of their actions



https://www.buzzfeed.com/ashleyperez/creepers-r-us

Unintended Audiences



https://www.ted.com/talks/alessandro_acquisti_why_privacy_matters#t-53301

University of Kent

Computing

Top Concerns

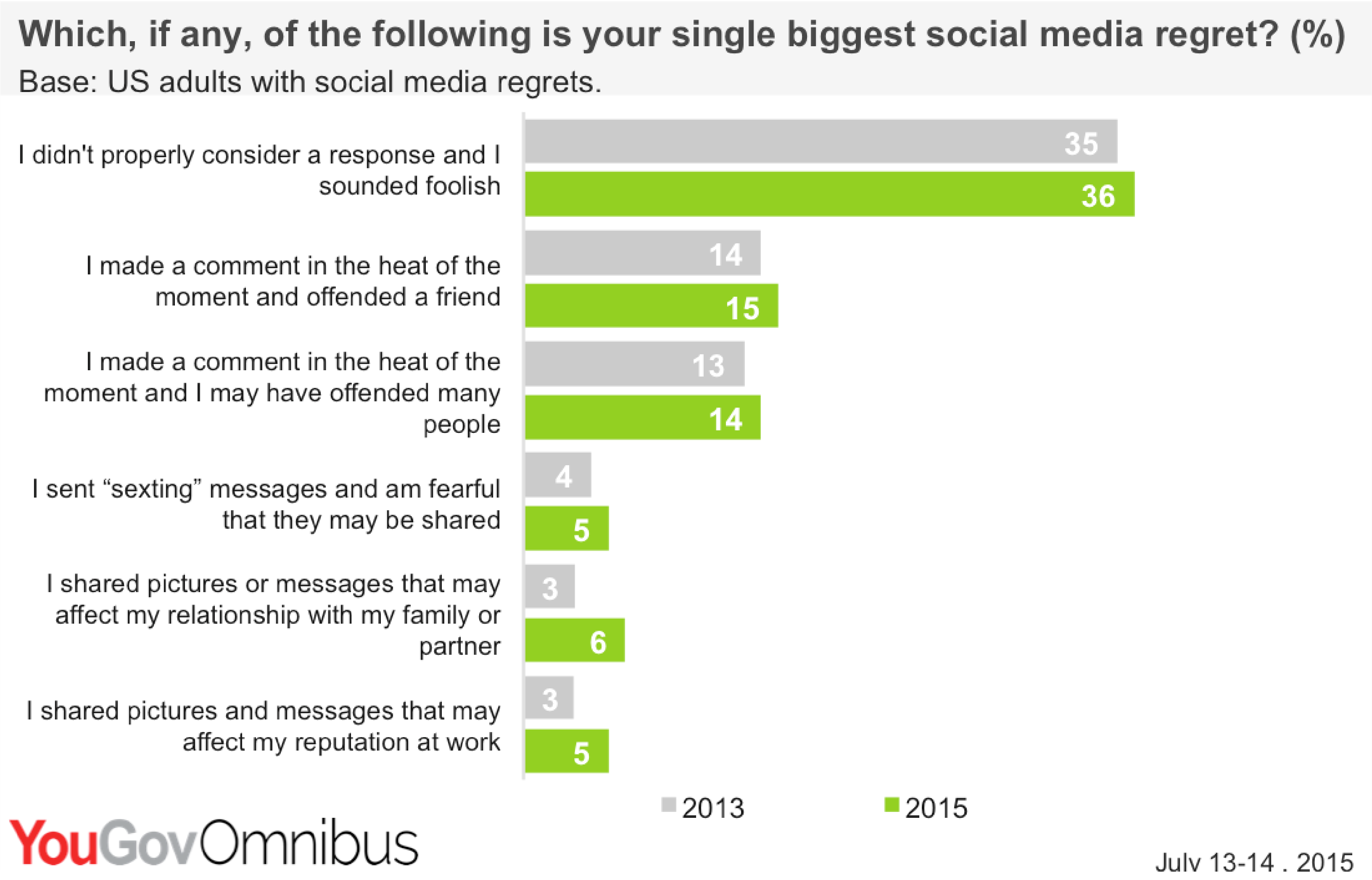
Scenario	Concerned
1. Thieves using Facebook to track, monitor, locate, and identify you as a potential victim.	68.8%
2. Your employer seeing an inappropriate photo or comment on your profile.	62.7%
3. Your employer using your profile to assess your suitability for the company.	55.0%
4. Sexual predators using Facebook to track, monitor, locate, and identify you as a potential victim.	51.9%
5. Your employer using Facebook to monitor your conduct while you're at work.	46.2%
6. Your employer using Facebook to monitor your conduct while you're away from work.	44.6%
7. A stranger will see an inappropriate photo or comment on your profile.	40.8%
8. Political parties using Facebook to target you through the use of ads and data mining.	30.4%
9. Your university using Facebook to identify you as a university code violator.	20.0%
10. Law enforcement using Facebook to track drug use and other illegal activities.	17.3%

Johnson et al. Facebook and Privacy: It's Complicated
Symposium on Usable Privacy and Security (SOUPS), pages 9:1–9:15, 2012

Mitigation Strategies

- Have friends only profiles, custom friend lists (subsets of friend network)
  - Not for privacy purposes though
  - Other features, e.g. group friends who play the same game
- Curating friend network: Deny friend request, unfriend
- Delete posts, untag themselves
- Go beyond official privacy controls provided by the OSNs
  - Multiple accounts: Maintain separate profiles, separate OSNs for different purposes
  - Ask friends to remove photos

Regrets



http://www.huffingtonpost.com/shane-paul-neil/more-than-half-of-america_b_7872514.html

University of Kent

Computing

Multi-party Privacy

Fogues et al. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. ACM Transactions on Computer-Human Interaction, 24(1):5:1-5:29, 2017

To Post or Not to Post?

Picture and Context

Relationship: Friends (98.3%)
Sensitivity rating: $\mu = 3.29$ ($\sigma = 1.16$)
Sentiment rating: $\mu = 3.82$ ($\sigma = 1.11$)

Description Three friends, Santosh, Arun, and Nitin, decided to perform some stunts on a motorcycle. Unfortunately, while performing a stunt, Arun and Nitin had a minor accident. Santosh took the picture below at that very moment. Santosh wants to upload the picture to his social media account.

Arguments

Positive consequence argument. Fortunately, none of us got hurt. This picture makes anyone who sees it laugh out loud.
Negative consequence argument. People looking at this picture may think that we are reckless drivers, which is not true.
Exceptional case argument. Motorbike stunts are not something we do every-day.

Picture and Context

Relationship: Colleagues (92.9%)
Sensitivity rating: $\mu = 3.26$ ($\sigma = 1.41$)
Sentiment rating: $\mu = 2.46$ ($\sigma = 1.50$)

Description Jerry, Laura, and Sabrina work together in a company. They were asked to attend the Christmas party dressed. However, a guy in their company (the one in pink dress) brought the whole dressing to a new level. They took the following picture at the party. Jerry wants to upload the picture to his social media account, a few days after the party.

Fogues et al. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. ACM Transactions on Computer-Human Interaction, 24(1):5:1-5:29, 2017

Anonymisation of Datasets

- Data owner, e.g. hospital
- Has private dataset with user specific data
- Goal: To share a version of the dataset with researchers
  - Dataset can help researchers to train better models
  - Results can help the data owner
- Provide scientific guarantees that users in the dataset cannot be re-identified
- Data should remain practically useful

Medical Data

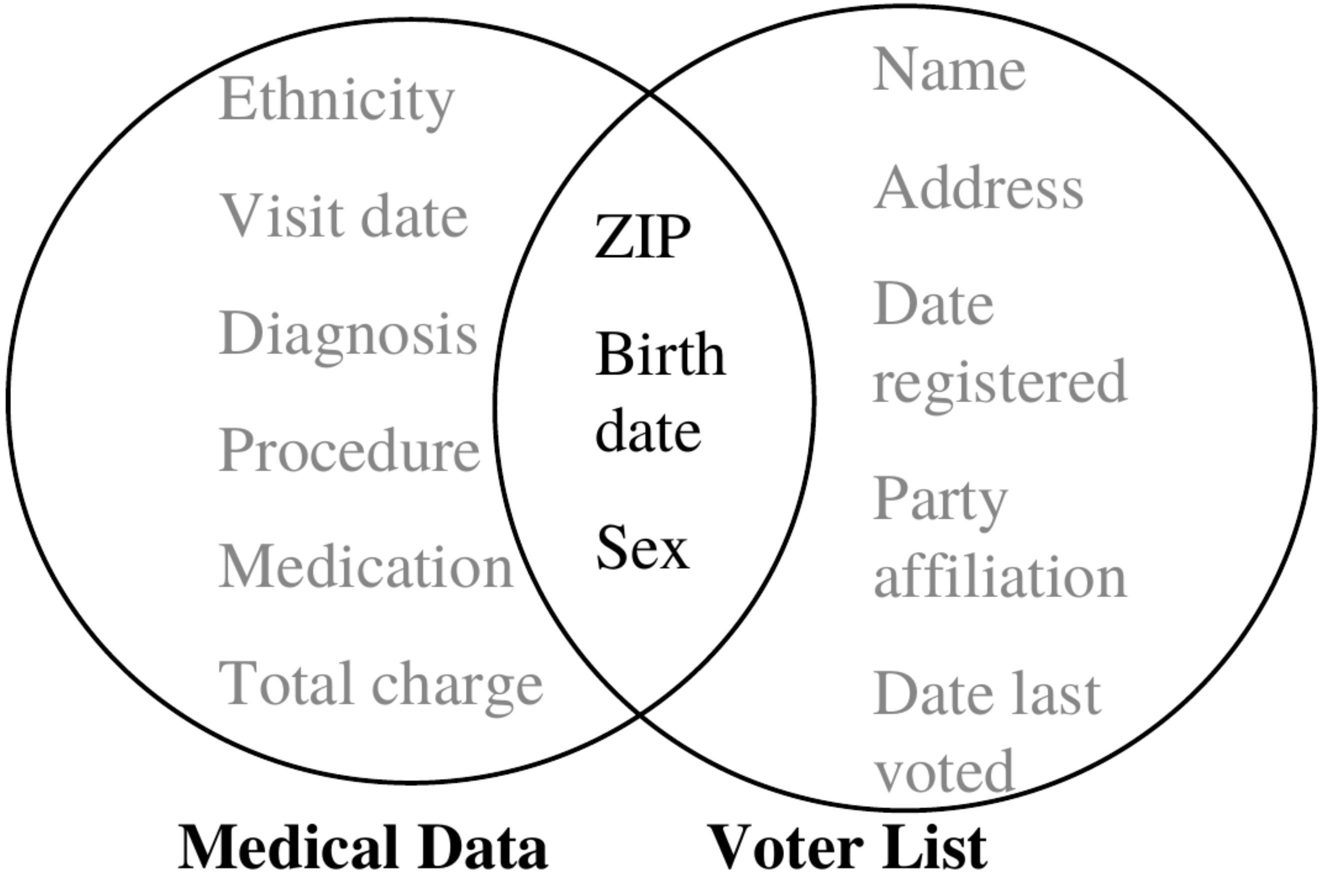
	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	< 30	*	AIDS
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Real Problem

- 87% (216M of 248M) of the US population
- Uniquely identifiable based only on
  - 5-digit ZIP code
  - Gender
  - Date of birth

Sweeney. Uniqueness of Simple Demographics in the US Population, 2000

Re-identification by Linking



Re-identification of Individuals

- William Weld: Governor of MA at the time
- His medical record in the Group Insurance Commission (GIC) data
- Lived in Cambridge, MA
- From the voter list
  - Six people with his particular birth date
  - Three of them male
  - He was the only one in his ZIP code

Quasi-identifiers

- Attributes that in combination can uniquely identify individuals
- Data owner should identify the quasi-identifier

Zip Code	Gender	Date of Birth	Medical Condition
**	**	**	**
**	**	**	**

`nonsensitive' (at least individually)

sensitive

Sensitive Columns

- Table with three columns
  - Doctor
  - Patient
  - Medication
- Which combinations are sensitive?
  - R(Doctor, Patient): Sensitive?
  - R(Doctor, Medication): Sensitive?
  - R(Patient, Medication): Sensitive?

Netflix Prize

- In October 2006, Netflix offered a $1M prize for a 10% improvement in its recommendation system
- Released a training dataset for competitors to train their systems
- Disclaimer: To protect customer privacy, all personal information identifying individual customers has been removed and all customer IDs have been replaced by randomly assigned IDs

Problems

- Netflix is not the only movie-rating portal on the web
- On IMDb, individuals can rate movies “not” anonymously
- Researchers from University of Texas at Austin linked Netflix dataset with IMDb to de-anonymise the identity of some users

De-anonymisation of Large Datasets

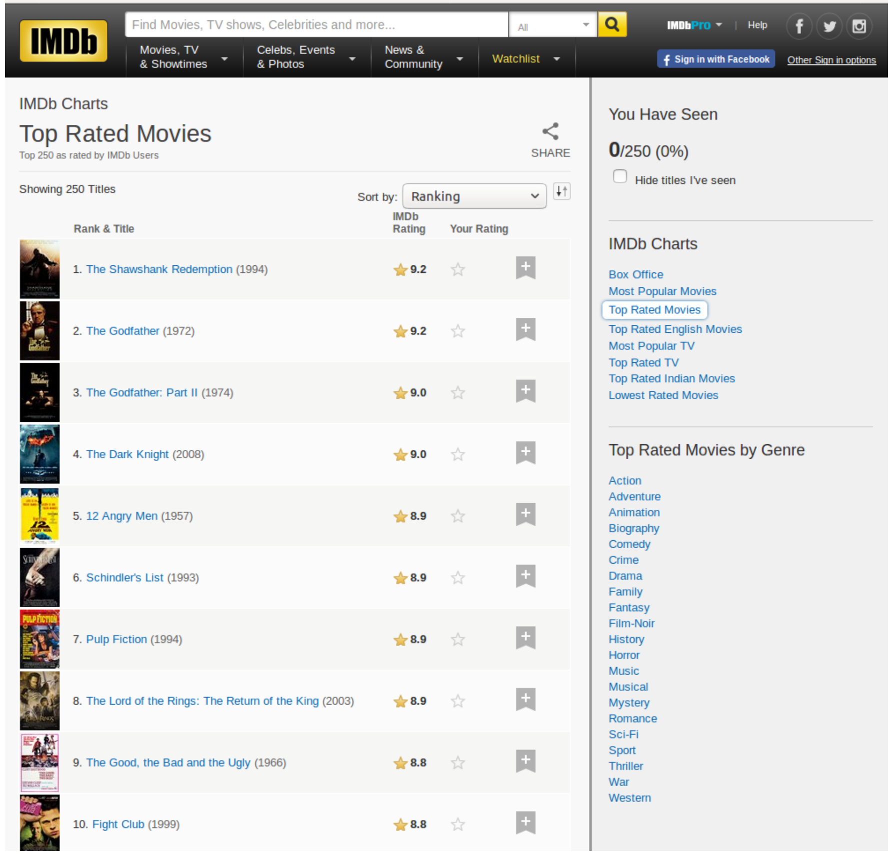
- De-anonymisation attacks
- Linking datasets (public or private) together to gain additional information about users
- Even if sensitive attributes are not contained in the dataset, they can be inferred with high accuracy

Narayanan and Shmatikov. Robust De-anonymization of Large Sparse Datasets
IEEE Symposium on Security and Privacy, pages 111-125, 2008

Netflix Dataset

- “Anonymous” movie ratings of 480,189 subscribers of Netflix
- 100,480,507 movie ratings
- Between 1999 and 2005
- Less than 1/10 of the entire 2005 database

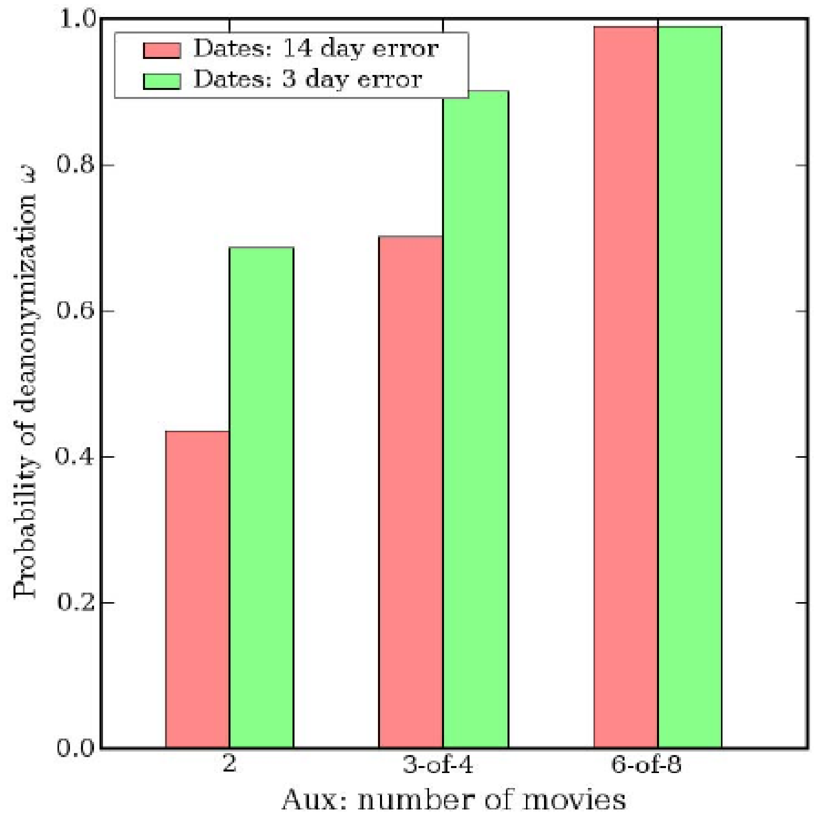
Public IMDb Ratings

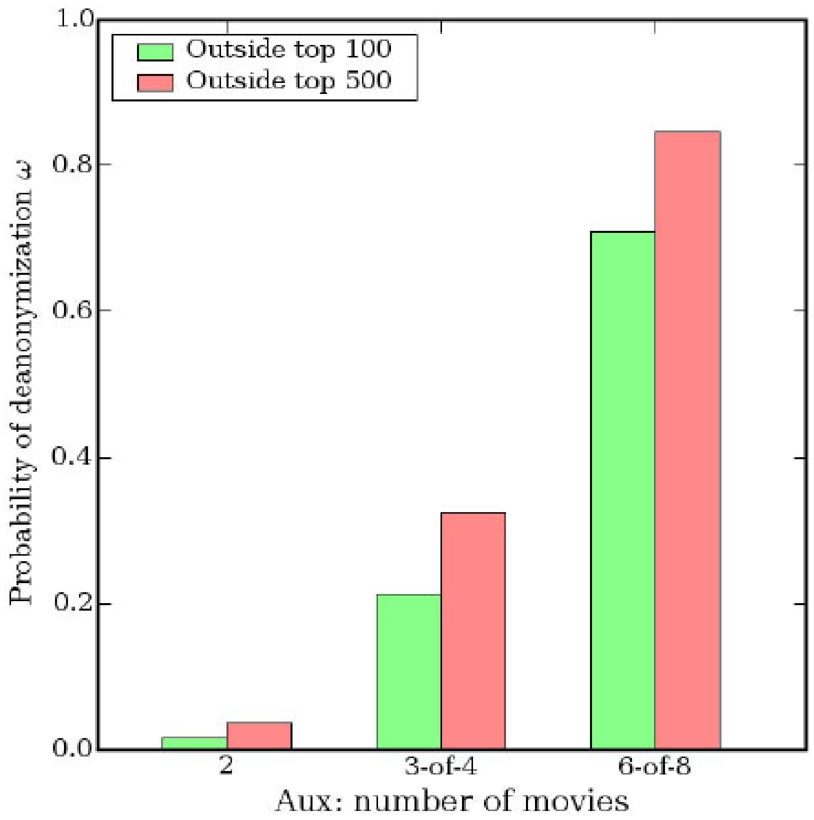


Results

- With 8 movie ratings known (2 of them might be completely wrong)
- And, dates having a 14-day error margin
- 99% of users can be uniquely identified
- With 2 ratings and 3-day error dates, 68% of users can be uniquely identified

De-identification Probability





Implications

- Why would someone who (not anonymously) rates movies on IMDb care about privacy of Netflix ratings?
  - Extract entire movie viewing history from Netflix
  - Infer political orientation
  - Infer religious views

Harvard's Privacy Meltdown

- In 2006, when Facebook was starting to emerge
- Harvard student Facebook data shared for research purposes
- Soon after release, some students were successfully re-identified

http://chronicle.com/article/Harvards-Privacy-Meltdown/128166/

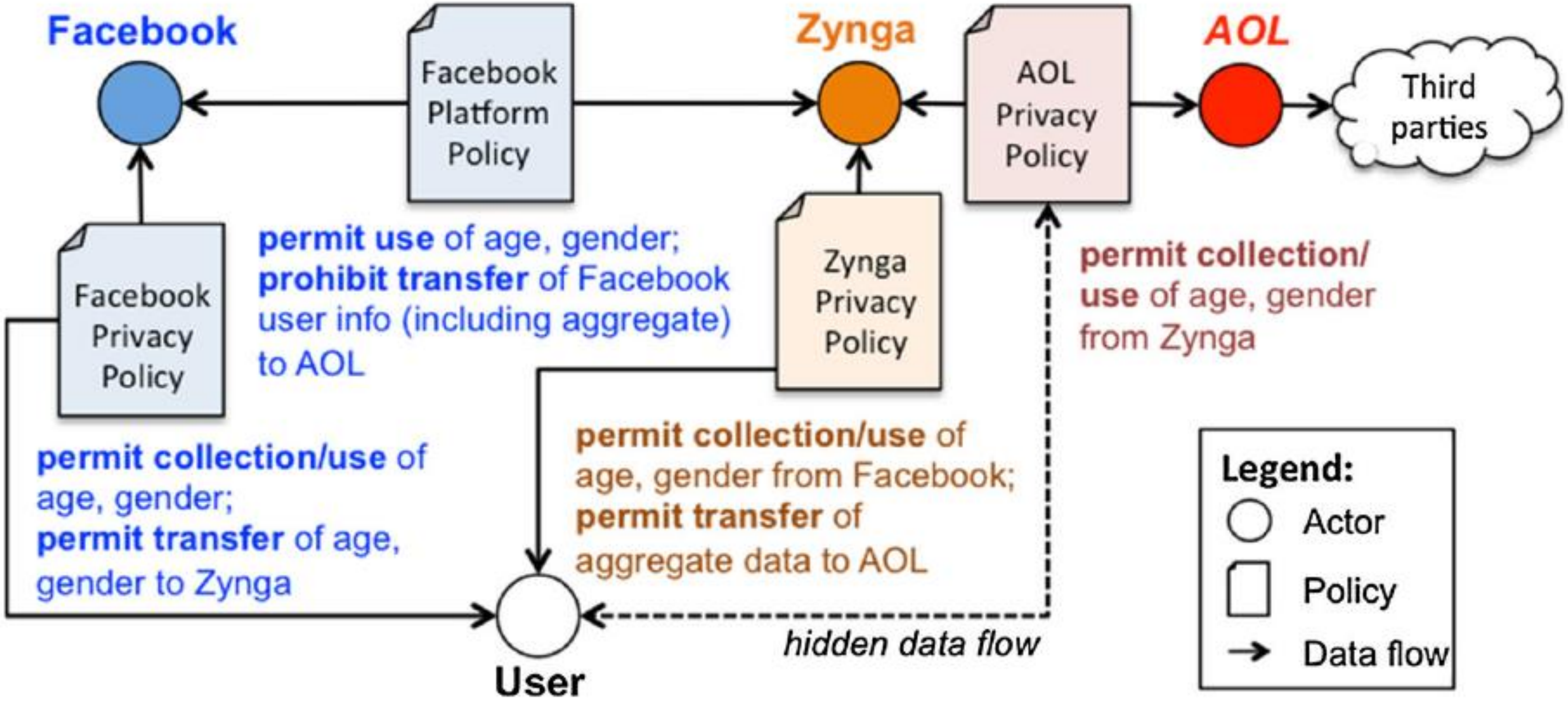
University of Kent

Computing

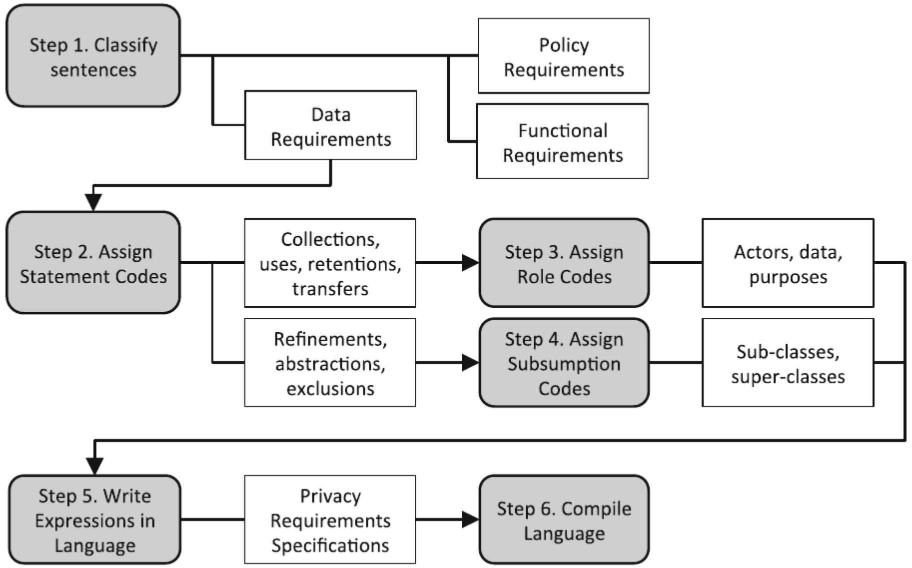
Facebook and Zynga

Breaux et al. Eddy, a Formal Language for Specifying and Analyzing Data Flow Specifications for Conflicting Privacy Requirements. Requirements Engineering, 19(3):281-307, 2014

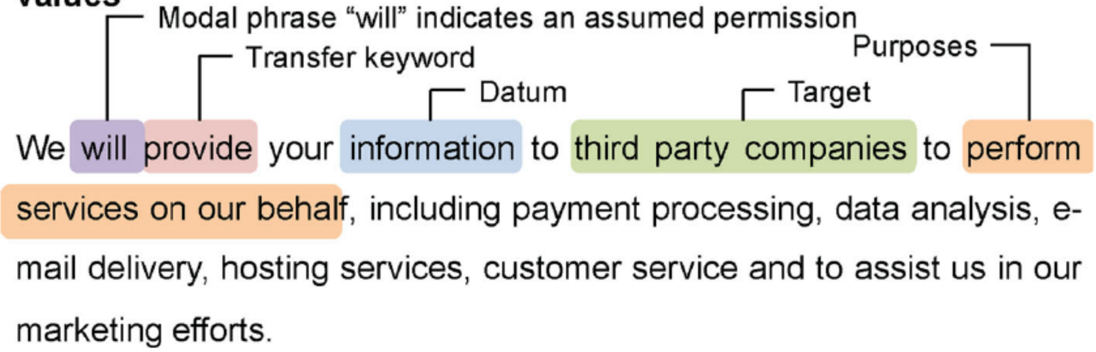
Conflicting Privacy Policies



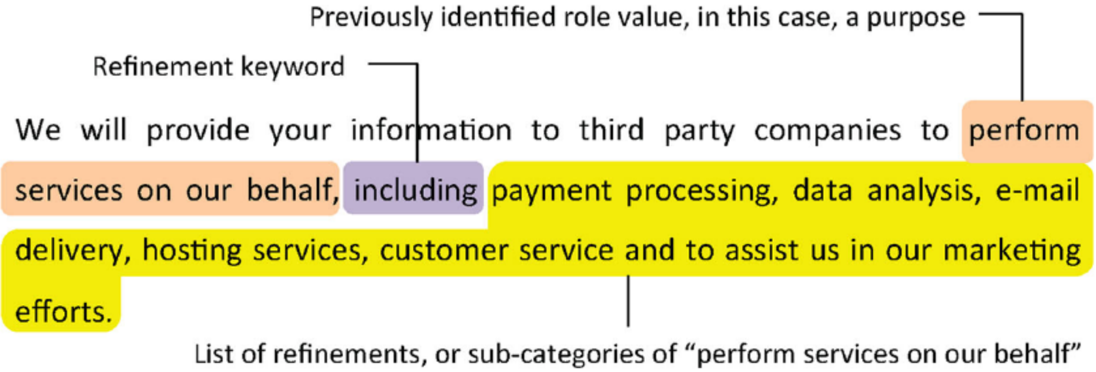
Policy Analysis



Step 3: Annotate policy text to identify action and role values



Step 4: Annotate policy text to identify other subsumption relations



OSN: The Legal Side

- Uniqueness:
  - Immediate (no validation)
  - Stays on record
  - Outreach

Organisational Use

- Directly engage with customers
- Receive feedback
- Targeted advertising

Legal Issues

- Retweeting false reports
- Unfair trading
- Data controller: Liability for user content
- Discrimination based on social media vetting
- Cyber bullying and harassment
- Data protection laws

University of Kent | Computing

Conclusions

- In this lecture, we have
  - Described what social computing is
  - Defined an OSN as a directed graph
  - Reviewed problems involving social networks
  - Seen why data anonymity is important and ways of protecting the privacy of users
  - Reviewed legal issues involving social networks

University of Kent | Computing

Additional Material

- Mislove et al. You Are Who You Know: Inferring User Profiles in Online Social Networks. Conference on Web Search and Data Mining, pages 251–260, 2010
- Harvard Facebook incident: https://www.chronicle.com/article/Harvards-Privacy-Meltdown/128166/
- In a Mood? Call Center Agents Can Tell: http://www.nytimes.com/2013/10/13/business/in-a-mood-call-center-agents-can-tell.html
- TED talk:
  - https://www.ted.com/talks/alessandro_acquisti_why_privacy_matters#t-53301