

Privacy in the Internet of the Future

Michael Backes

CISPA, Saarland University & MPI for Software Systems

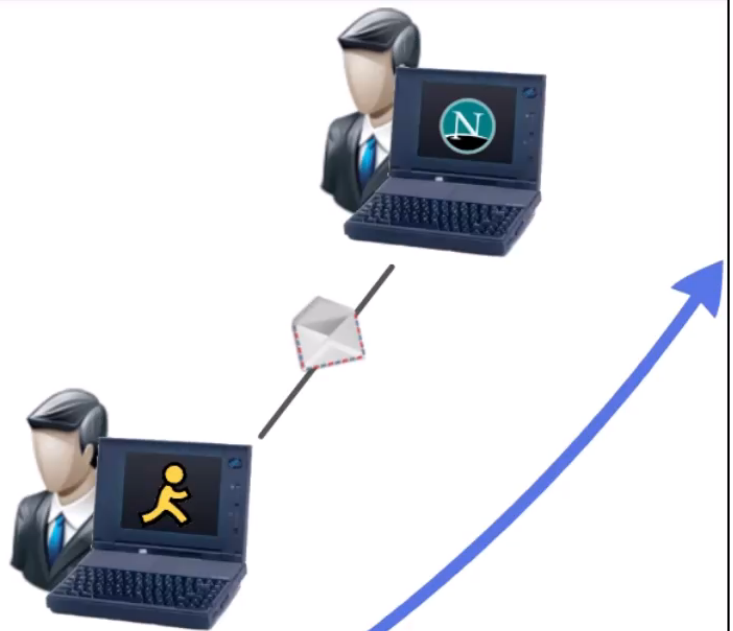
1990: Slow Internet

Primarily e-mail traffic



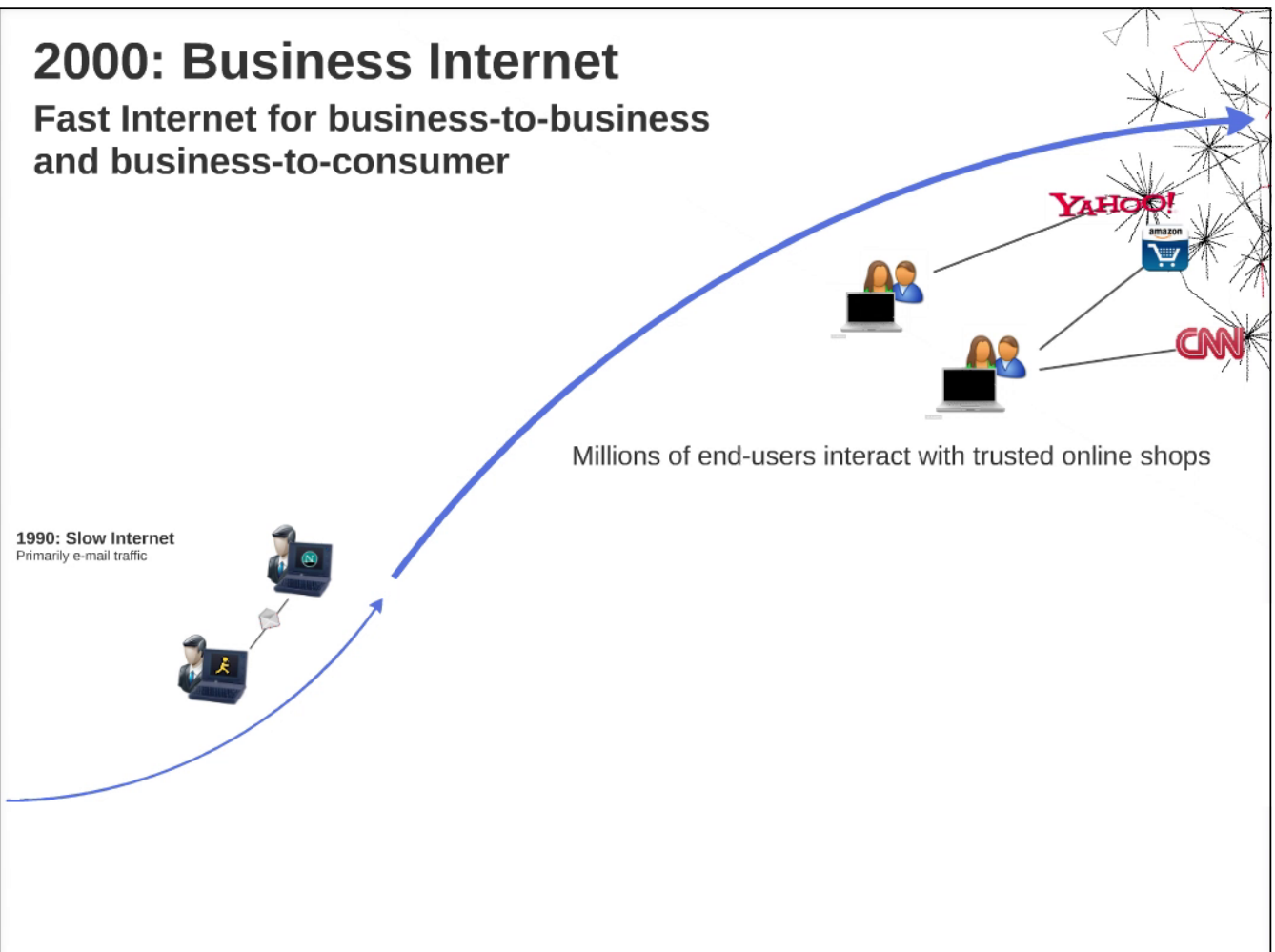
1990: Slow Internet

Primarily e-mail traffic



2000: Business Internet

Fast Internet for business-to-business and business-to-consumer



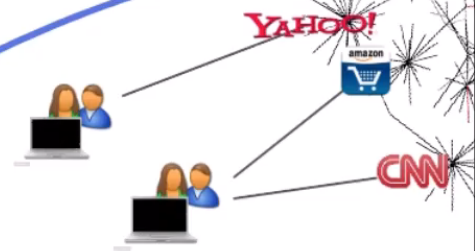
2000: Business Internet

Fast Internet for business-to-business and business-to-consumer

1990: Slow Internet
Primarily e-mail traffic



Millions of end-users interact with trusted online shops



Assumptions at that time:

- 1) Trusted computer connects to trusted services
- 2) Limited & controlled disclosure of personal information

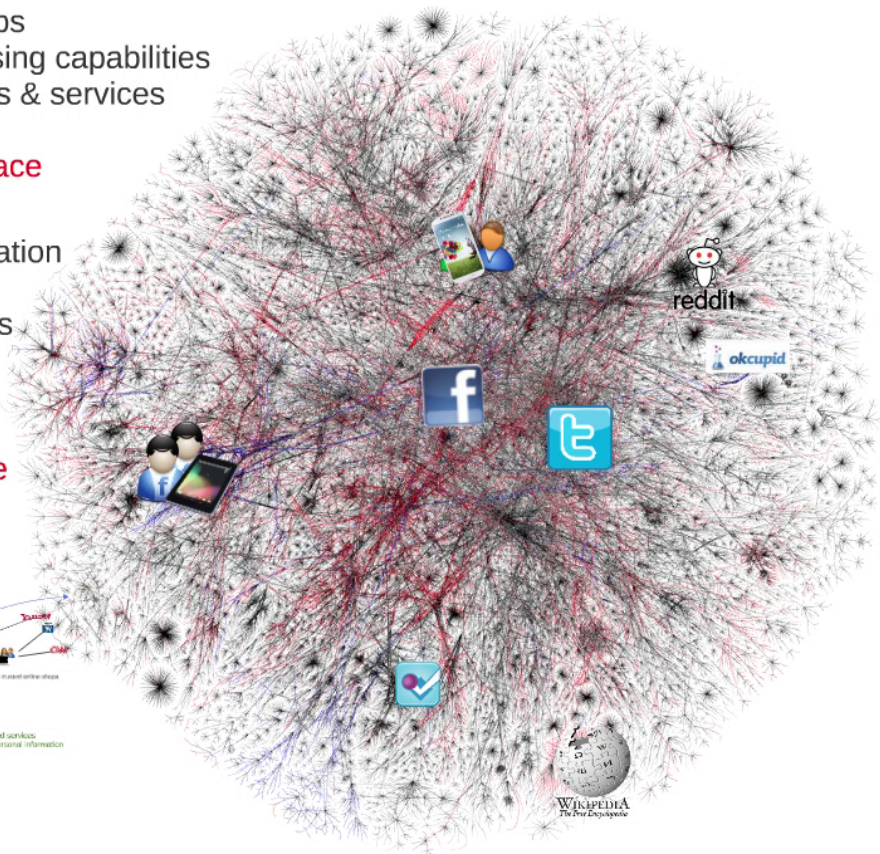
2015: User-Centric Internet

- 1) Complex trust relationships
 - Mobile devices with sensing capabilities
 - Third party software/apps & services

Vastly increased attack surface

- 2) Unprecedented dissemination of personal information
 - Advent of social networks & digital capture
 - Targeted advertisement

A deluge of privacy-sensitive information collected



2000: Business Internet
Fast Internet for business-to-business and business-to-consumer



Motivation – What is privacy?

- Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively.
- When something is private to a person, it usually means that something is inherently special or sensitive to them.
- The domain of privacy partially overlaps security, which can include the concepts of appropriate use, as well as protection of information.

Wikipedia (2014)



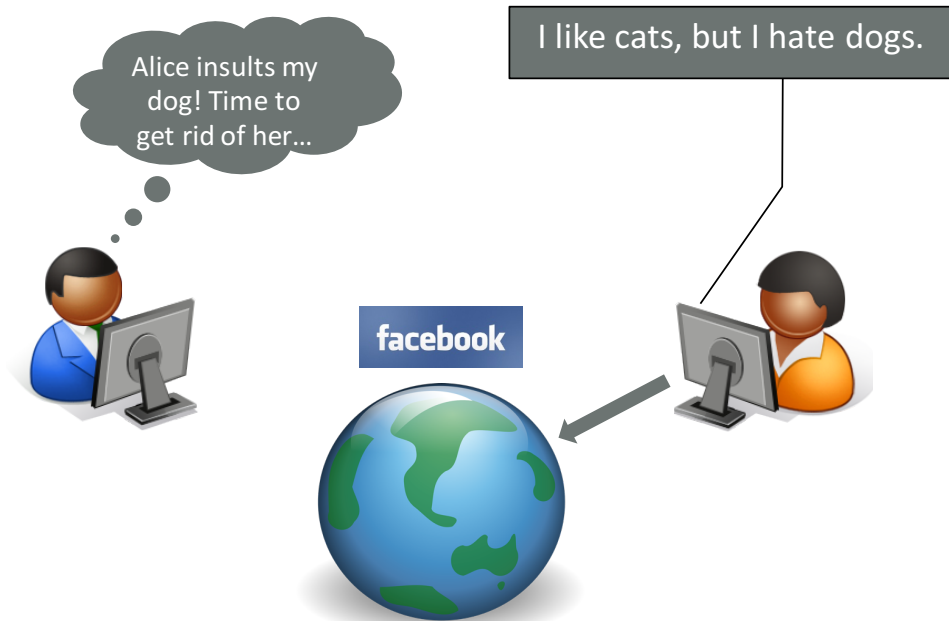
Motivation – Why do we need privacy?

- Sometimes, we do not want others to know something about us.



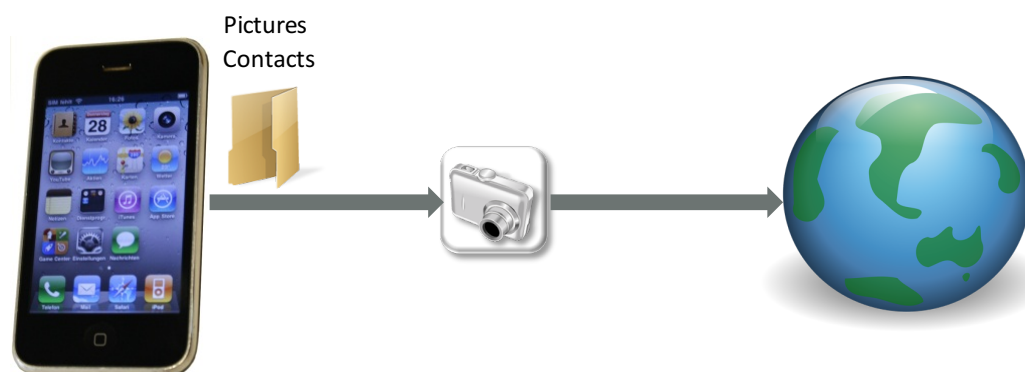
Motivation – Privacy in the Internet

- Alice shares her opinion in an Online Social Network.
- As a consequence, her employer, who dislikes that opinion, fires Alice.



Motivation – Privacy in the Internet

- Smartphone Apps can misuse your data.
- E.g. sell your contacts to an advertisement company.



Definitions – Privacy Breach

- A *privacy breach* occurs when a piece of sensitive information about an individual is disclosed to an adversary, someone whose goal is to compromise privacy.

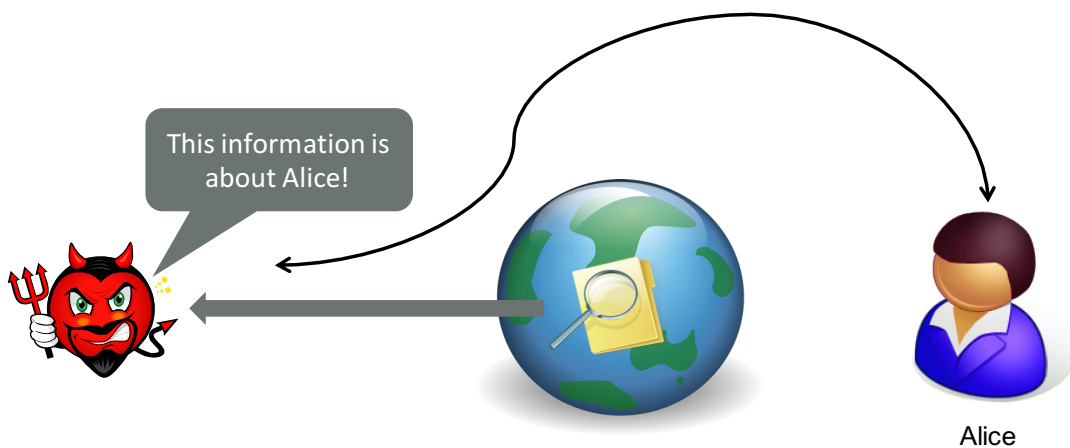
PRIVACY IN SOCIAL NETWORKS: A SURVEY (2011)



Definitions – Identity Disclosure

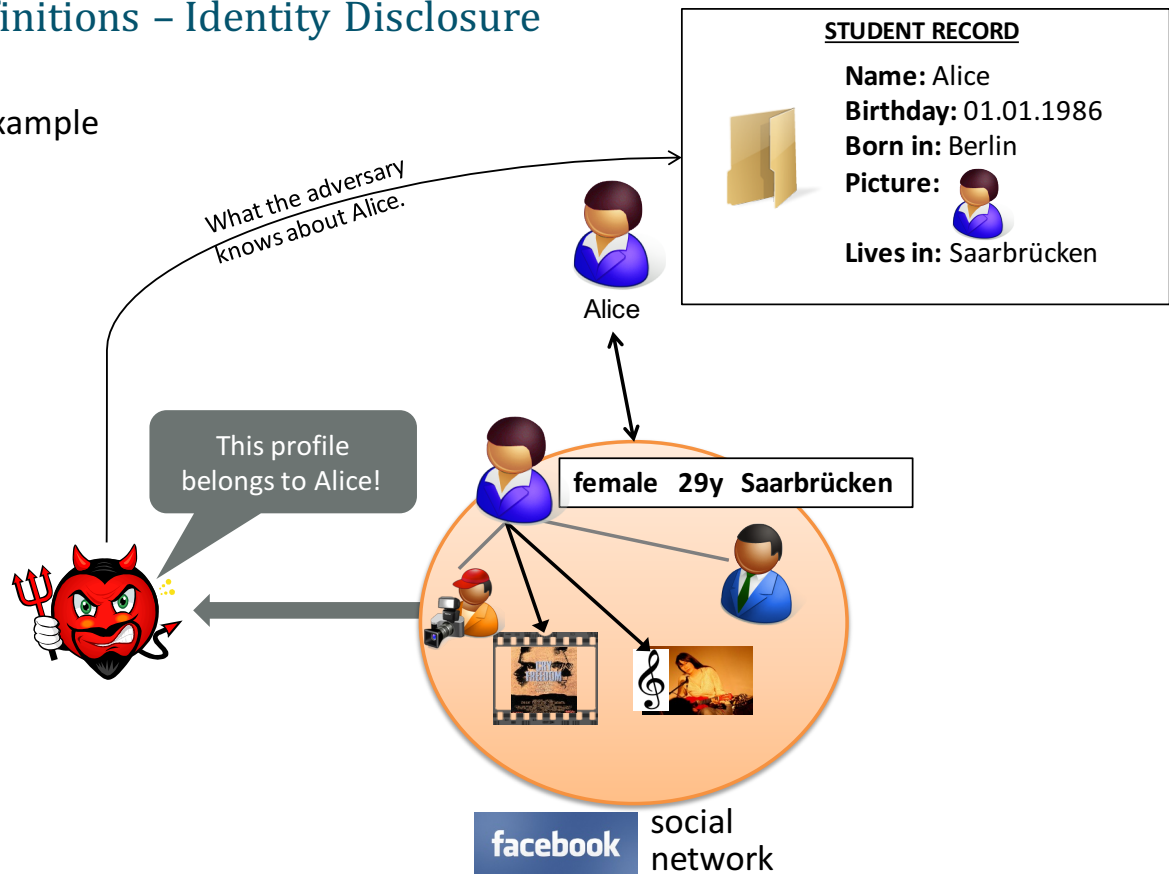
- *Identity disclosure* occurs when an adversary is able to determine the mapping from a profile v in the social network to a specific real-world entity p .

PRIVACY IN SOCIAL NETWORKS: A SURVEY (2011)



Definitions – Identity Disclosure

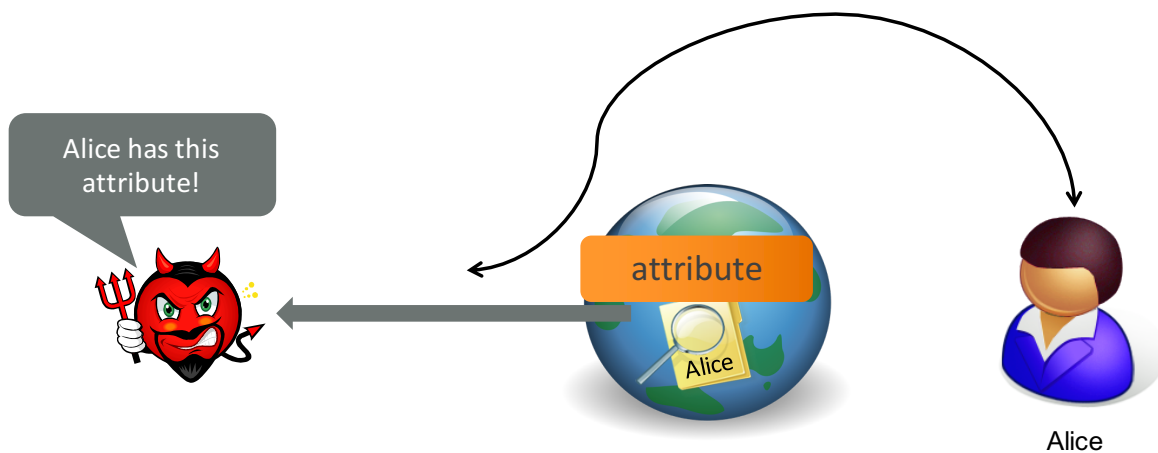
- Example



Definitions – Attribute Disclosure

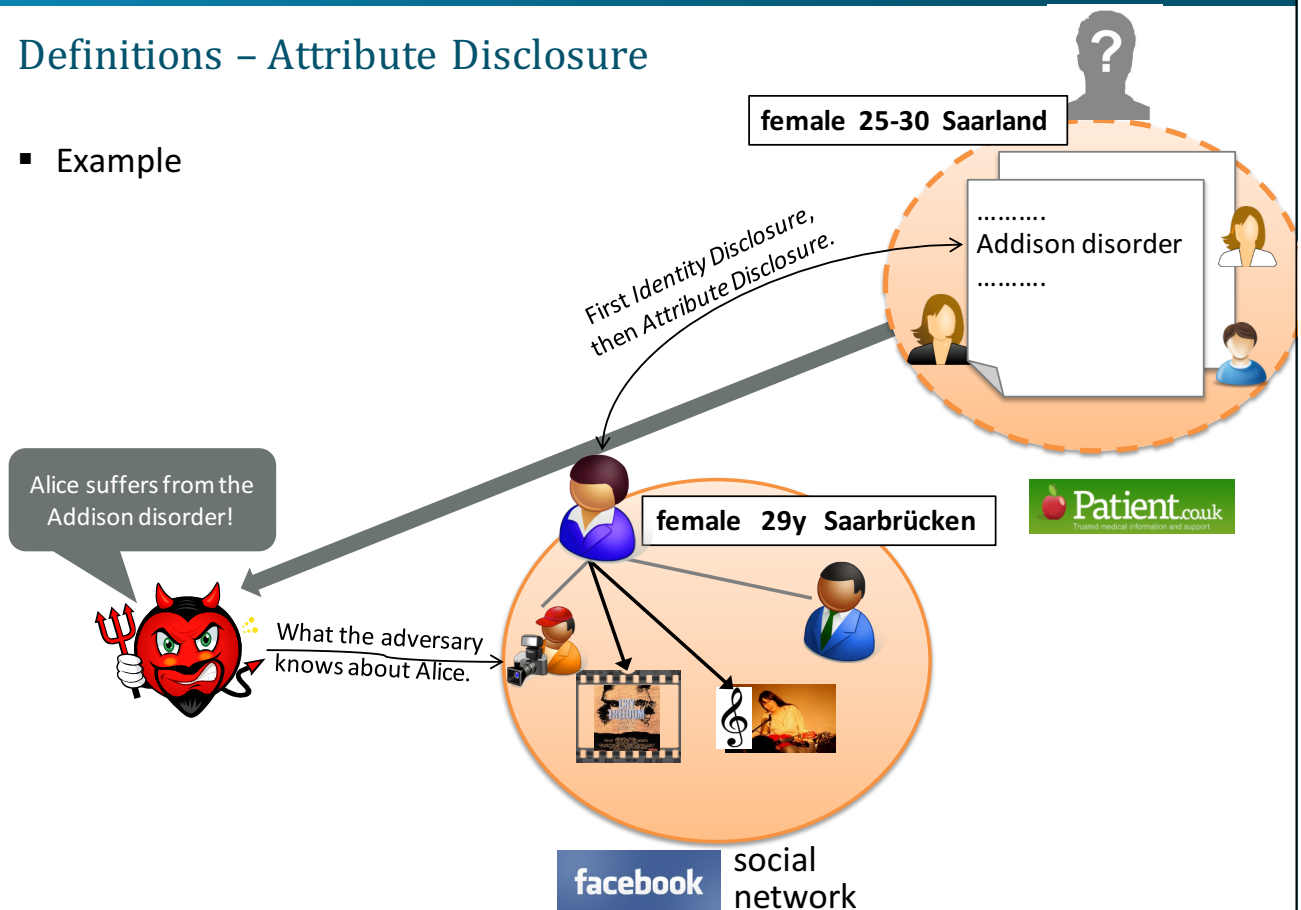
- Attribute disclosure* occurs when an adversary is able to determine the value of a sensitive user attribute, one that the user intended to stay private.

PRIVACY IN SOCIAL NETWORKS: A SURVEY (2011)



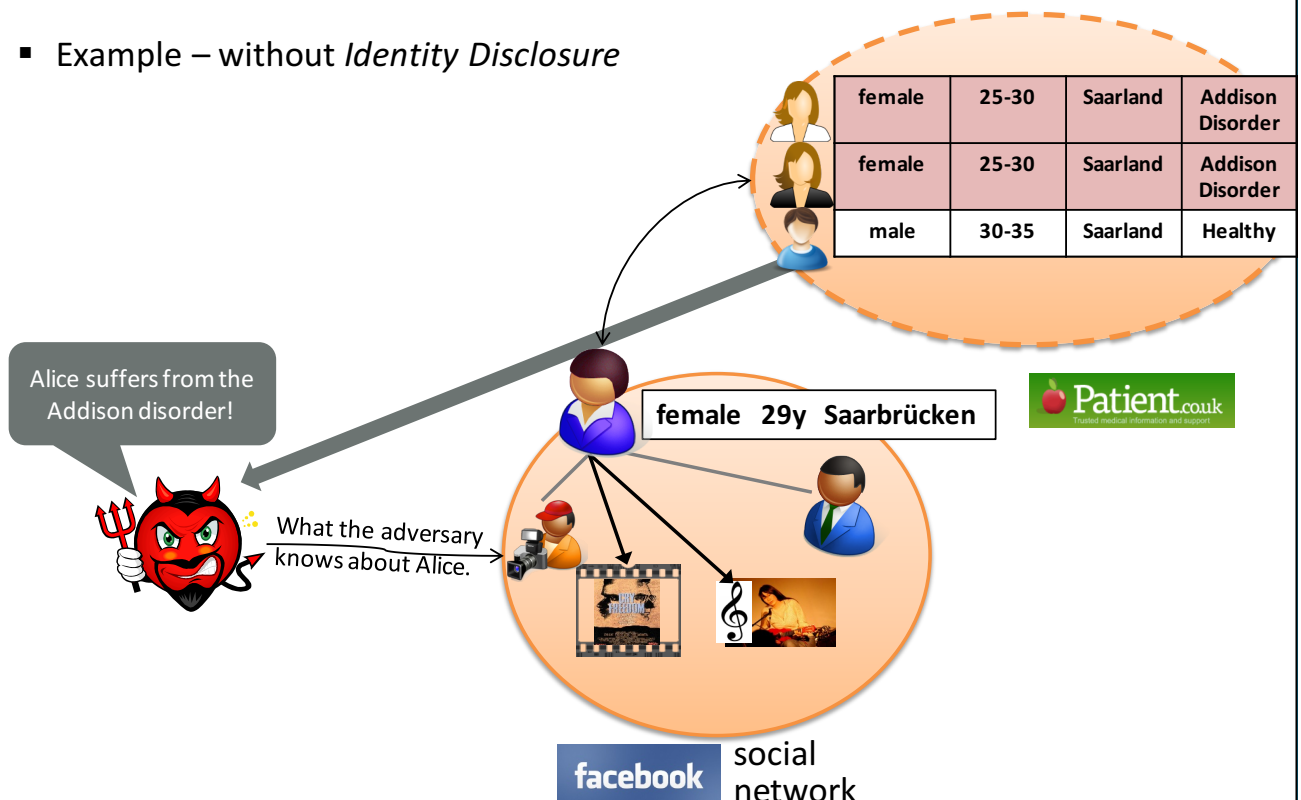
Definitions – Attribute Disclosure

- Example



Definitions – Attribute Disclosure

- Example – without *Identity Disclosure*



Database Privacy

e.g. Global Company Database
 Disproportionally Gender

ID	Zip	Age	Gender	Salary
24	61045	26	W	120000
34	67834	34	M	40000
28	12365	47	W	60000
56	24654	41	M	180000
97	98034	32	M	55000
102	12534	29	W	140000

- Structured Data
- Differentiation between Key- and Sensitive Attributes
- various privacy notions that guarantee some kind of privacy for the whole dataset
 - k-anonymity
 - l-diversity
 - t-closeness

- But which are the sensitive attributes?
- Is this the only data we can access?

-> Literature has shown that Privacy is much more diverse a problem (e.g. Netflix Challenge)

Netflix Challenge (Narayanan and Shmatikov, S&P08)

- Pre-defining a set of „sensitive“ attributes does not make sense!

Netflix Challenge:

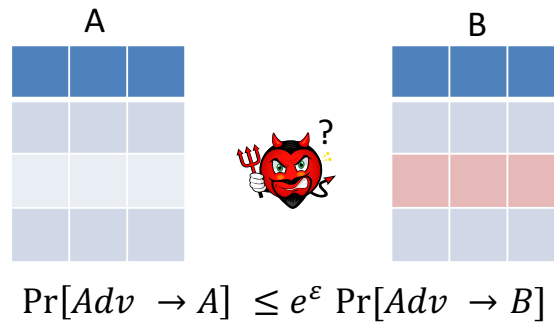
- Given: anonymized data-set of profiles with movie reviews
- Goal: identify anonymized profiles with live netflix data
- Use
 - movie reviews and
 - sparseness of data
 to identify profiles!

}

Movie reviews are sensitive data!
 Sparseness supplies auxiliary information!

Differential Privacy – for statistical databases

- Idea: Add noise to database to protect user data!



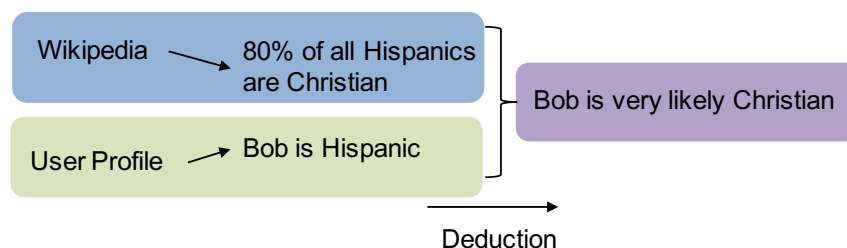
- Still restricted to structured and static data
- utility – privacy tradeoff unclear

Database Privacy vs Big-Data Privacy

Statistical Databases	Open Web/Big-Data
<ul style="list-style-type: none">▪ static and structured data▪ key- and sensitive attributes▪ mostly no adversarial background knowledge▪ privacy for the whole dataset	<ul style="list-style-type: none">▪ dynamic, heterogeneous and unstructured data▪ all information is potentially sensitive▪ ubiquitous background knowledge▪ whole dataset not known -> user centric privacy

Privacy in the open Web

- Evaluate the probability of unintended information leakage on the web
- Core Issues:
 - **Sensitivity of Information depends on context**
(e.g. discussing health issues in Facebook vs. on a health forum under an anonymous pseudonym)
 - **Unintended information disclosure through linked online profiles**
(e.g. linking anonymous profiles on various Forums to your Facebook account)
 - **Unintended information disclosure through inference**



Overview

This Lecture: **Identity Disclosure/Anonymity/Linkability** in the Internet

1. Network-level Anonymity

- Assessing Anonymity provided by Tor
- Monitoring Tor anonymity
- Network level and infrastructure adversaries

2. Semantic Linkability/Content-based Anonymity

- Assessing distinguishability of users by their content
- Linkability of anonymous profiles across communities
- Stylometry and assessing countermeasure effectiveness

Rigorously Assessing Anonymity of Tor

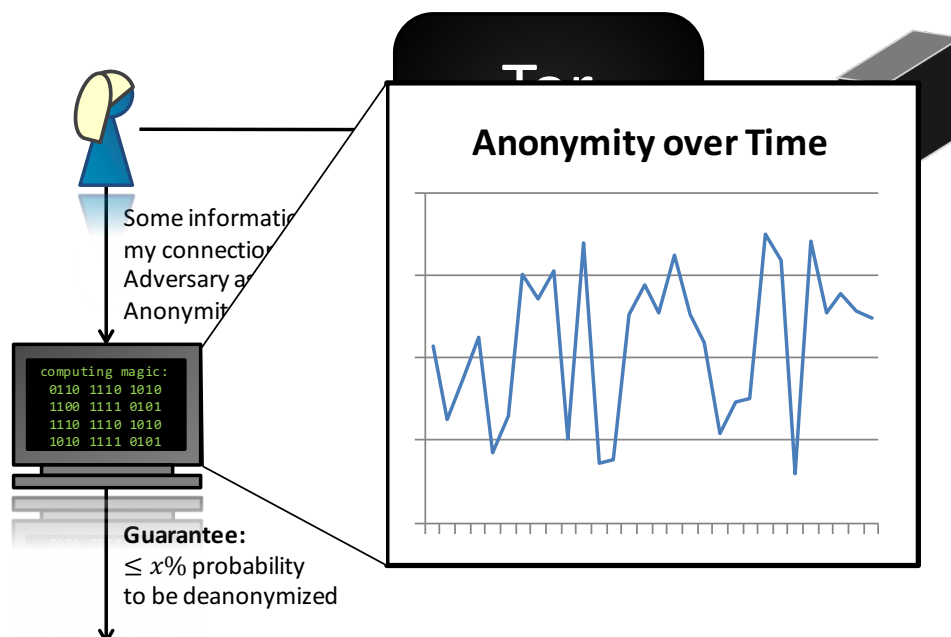
Michael Backes

(joint work with Simon Koch, Praveen Manoharan, Sebastian Meiser, Esfandiar Mohammadi, Marcin Slowik, Christian Rossow)

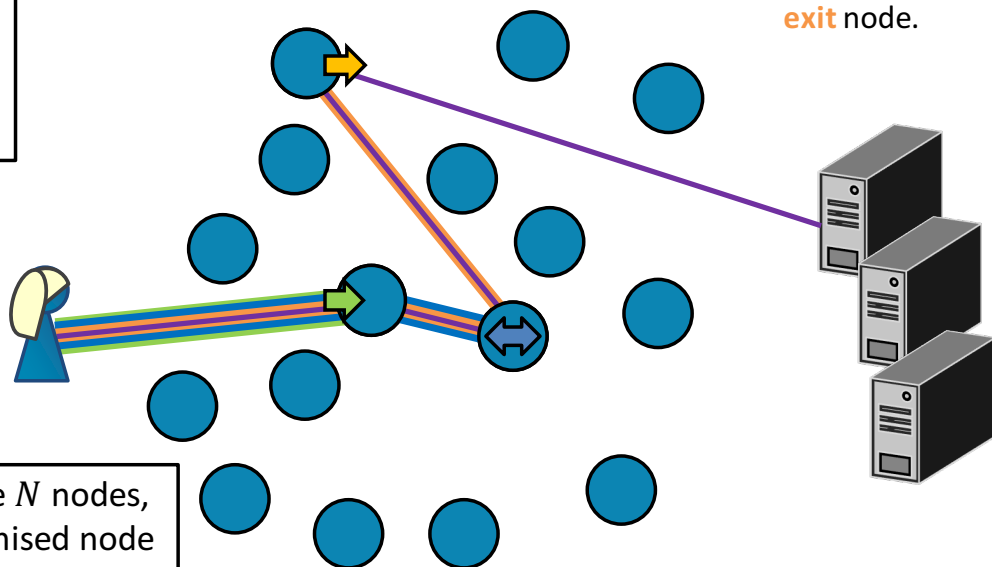
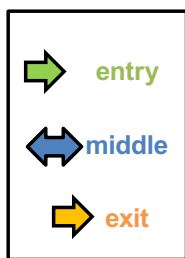
September 2, 2015

A Monitor for Anonymity (of Tor's path selection)

live



How Tor selects nodes (very general)



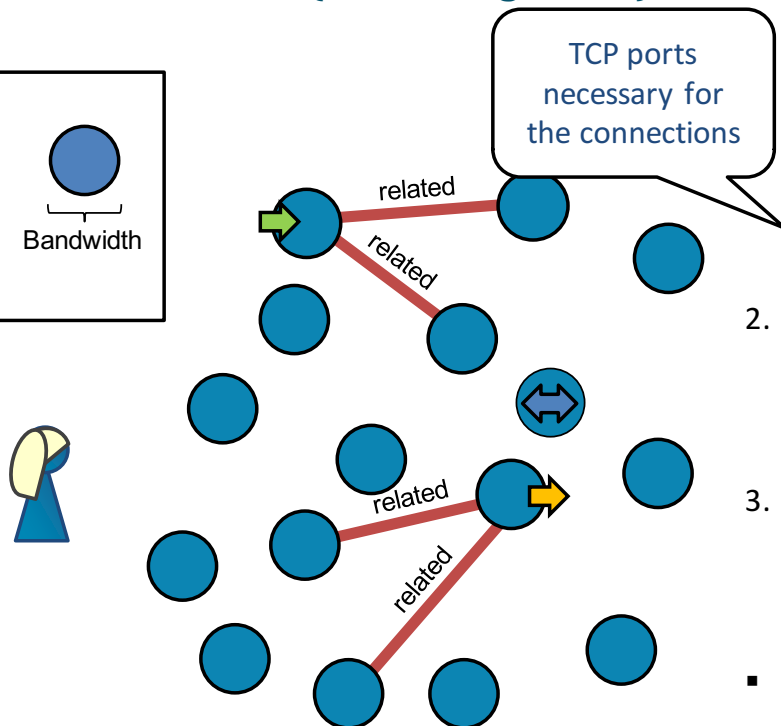
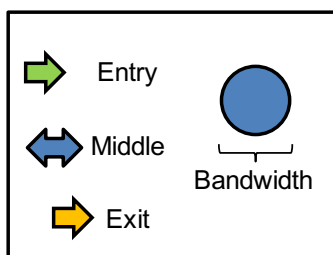
- Randomly choose an **entry**, a **middle** and an **exit** node.

If there are N nodes, a compromised node is chosen with probability $\frac{1}{N}$ as the **entry** node.

FOSAD 2015

24

How Tor selects nodes (a bit less general)

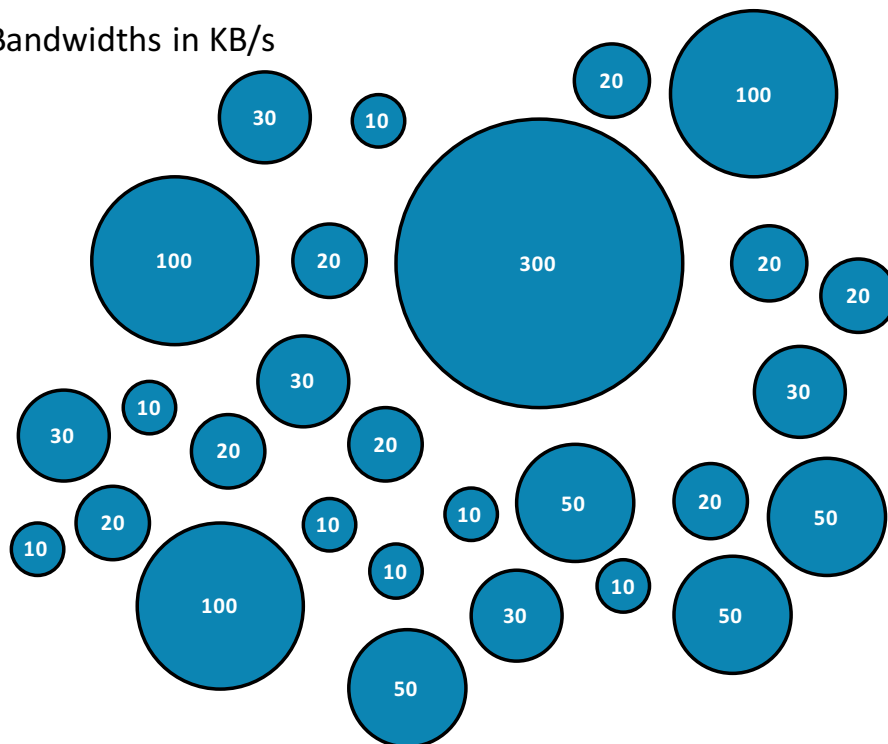


Randomly choose a possible **exit** node that allows Alice's **ports**, depending on its bandwidth.

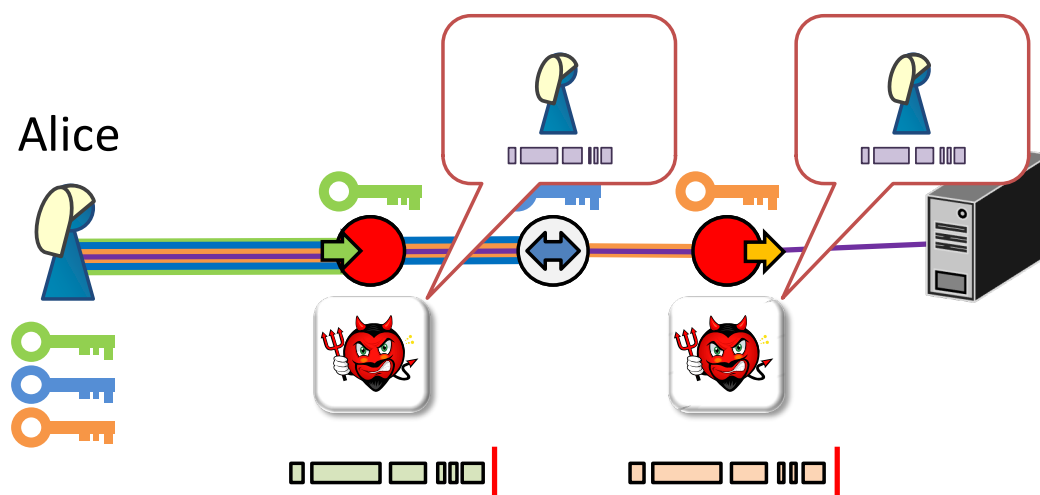
- Randomly choose a possible **entry** node, depending on its bandwidth
 - Randomly choose a possible **middle** node, depending on its bandwidth
- Never choose two **related** nodes in the same circuit.

(Imbalanced) Bandwidth of Tor Relays

- Bandwidths in KB/s

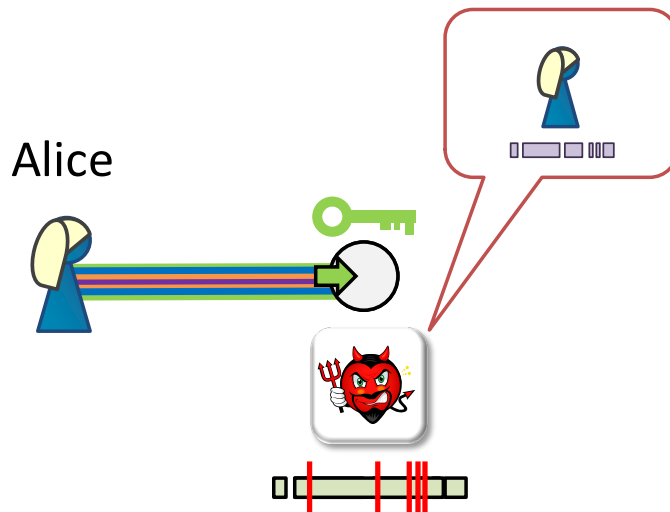


Effective attack: Traffic correlation



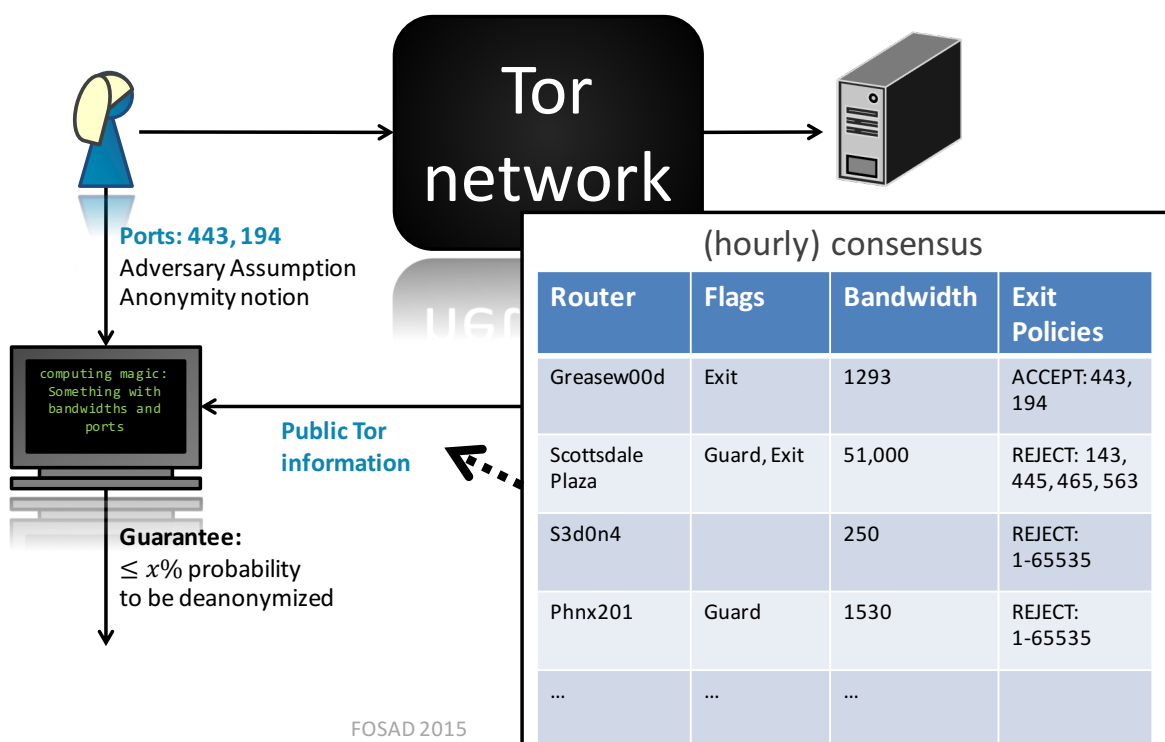
Although the traffic is encrypted and the encryptions look differently, the **pattern** remains.

Active Traffic correlation



Many countermeasures
(evening out traffic) can be
circumvented by active traffic
correlation attacks.

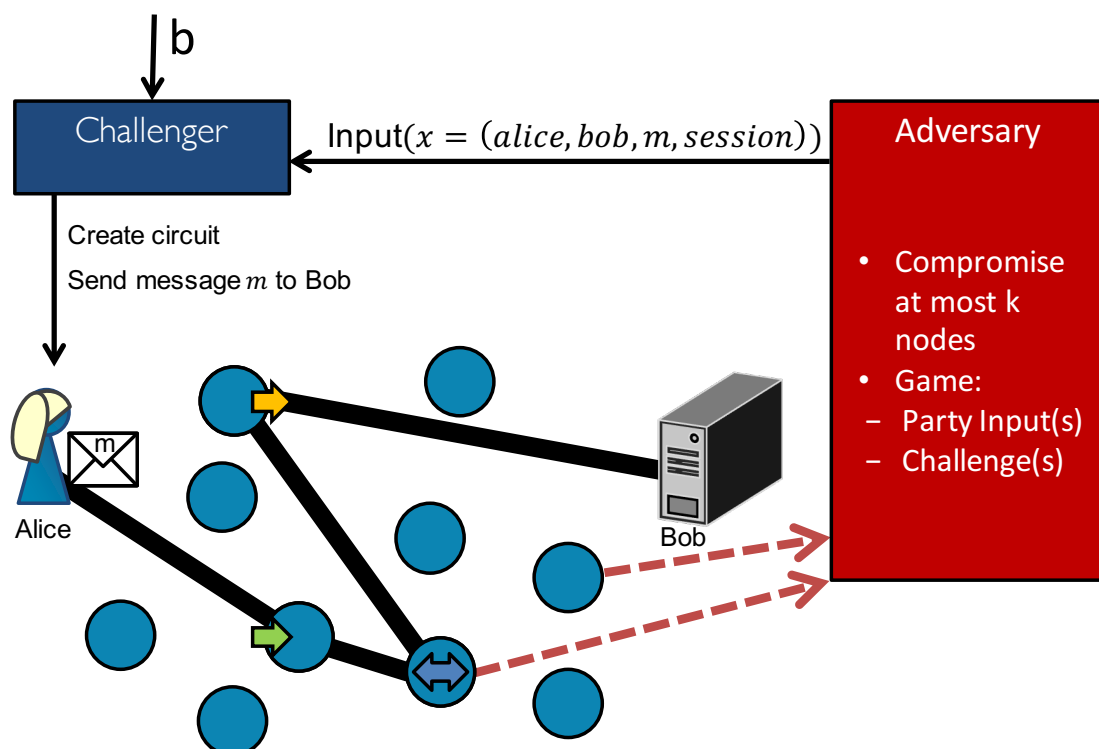
A live Monitor for Anonymity (of Tor's path selection)



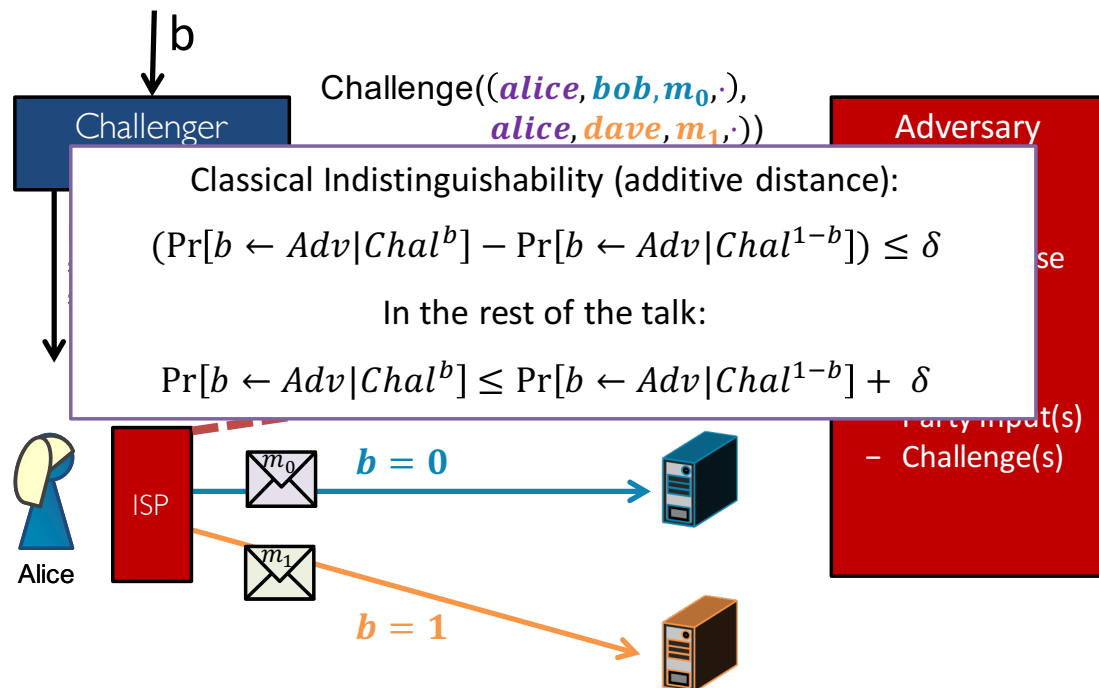
How to define anonymity?

(an indistinguishability game)

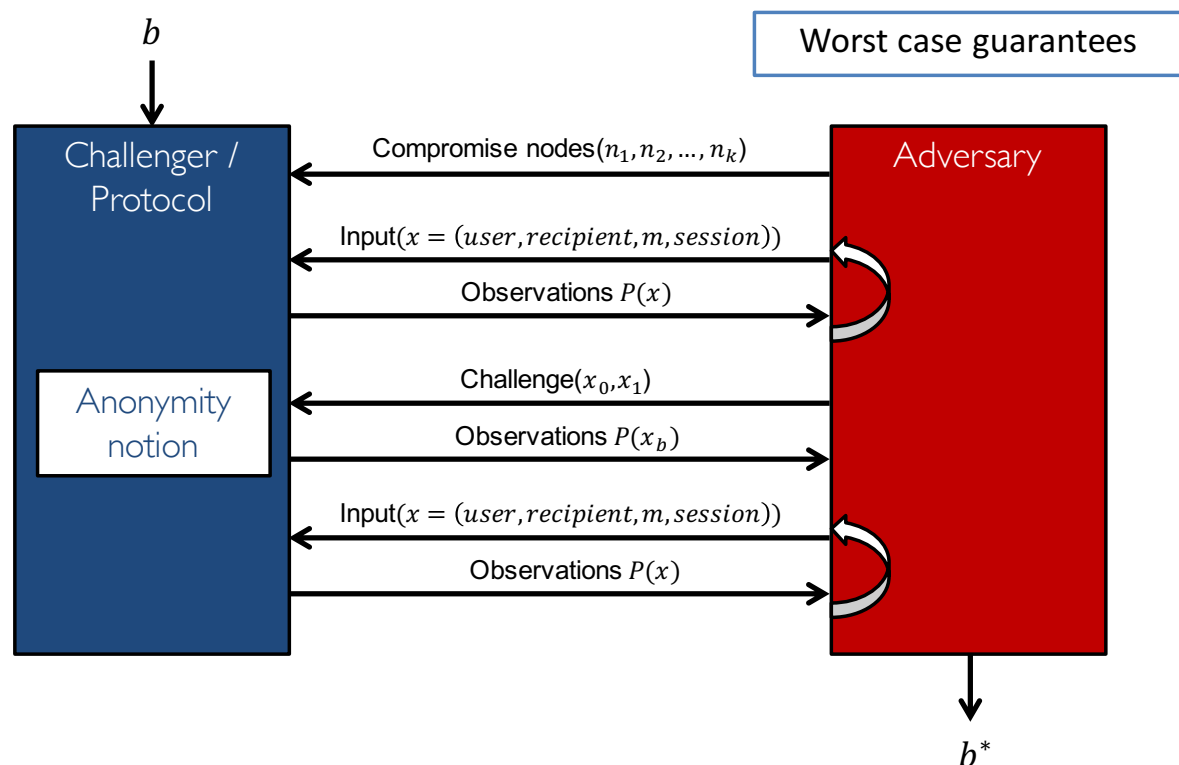
Defining Anonymity (1)



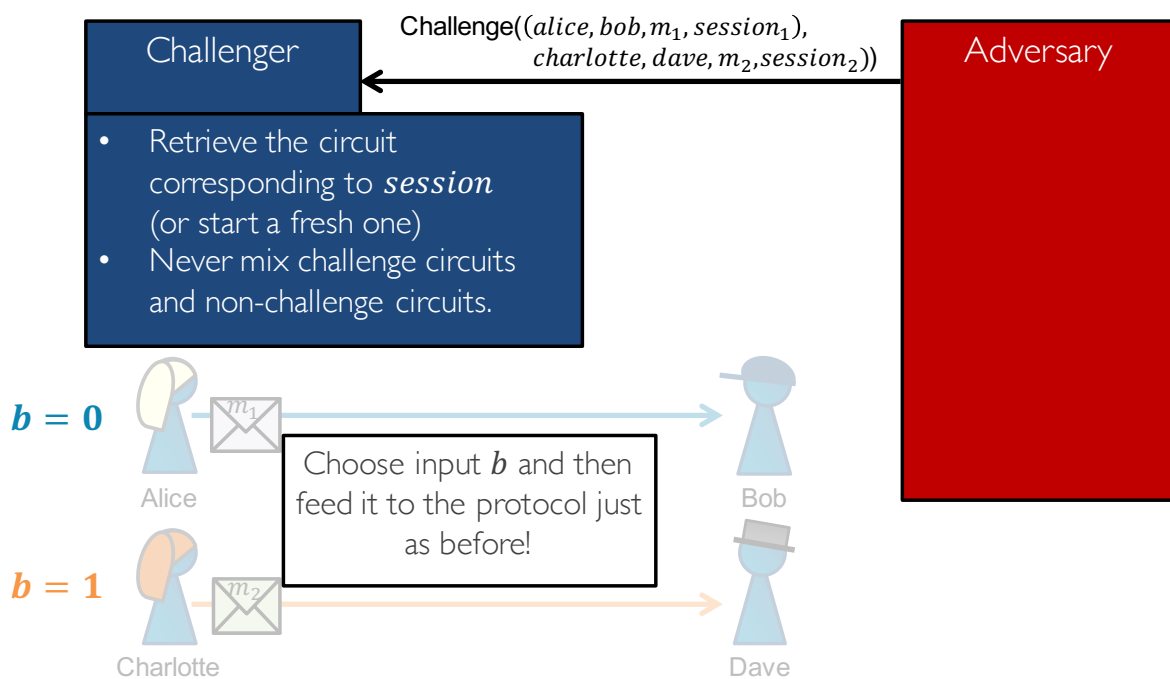
Defining Anonymity (2) – here: recipient anonymity



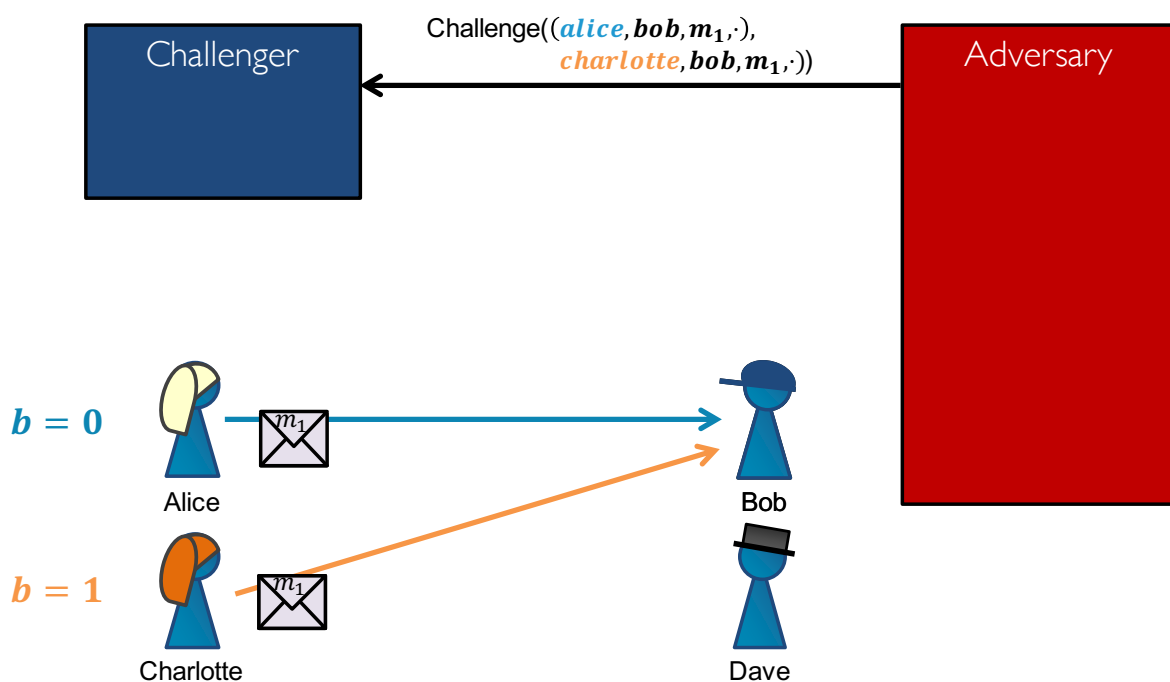
Quantifying Anonymity in AnoA



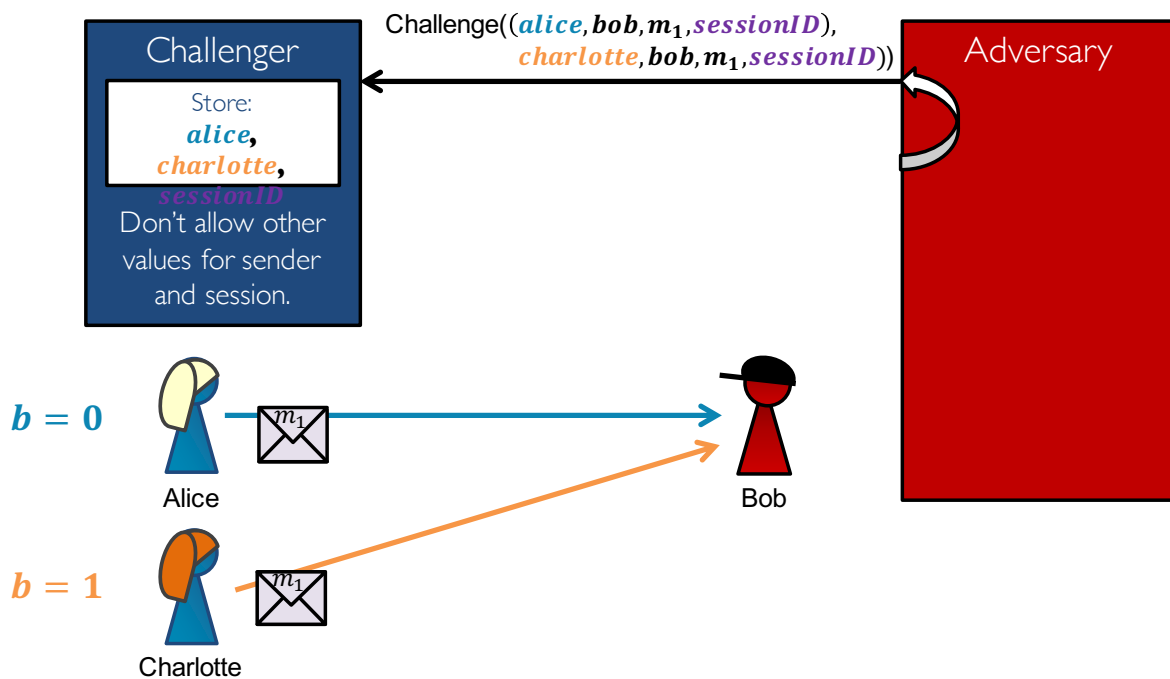
Challenger (with sessions) in detail



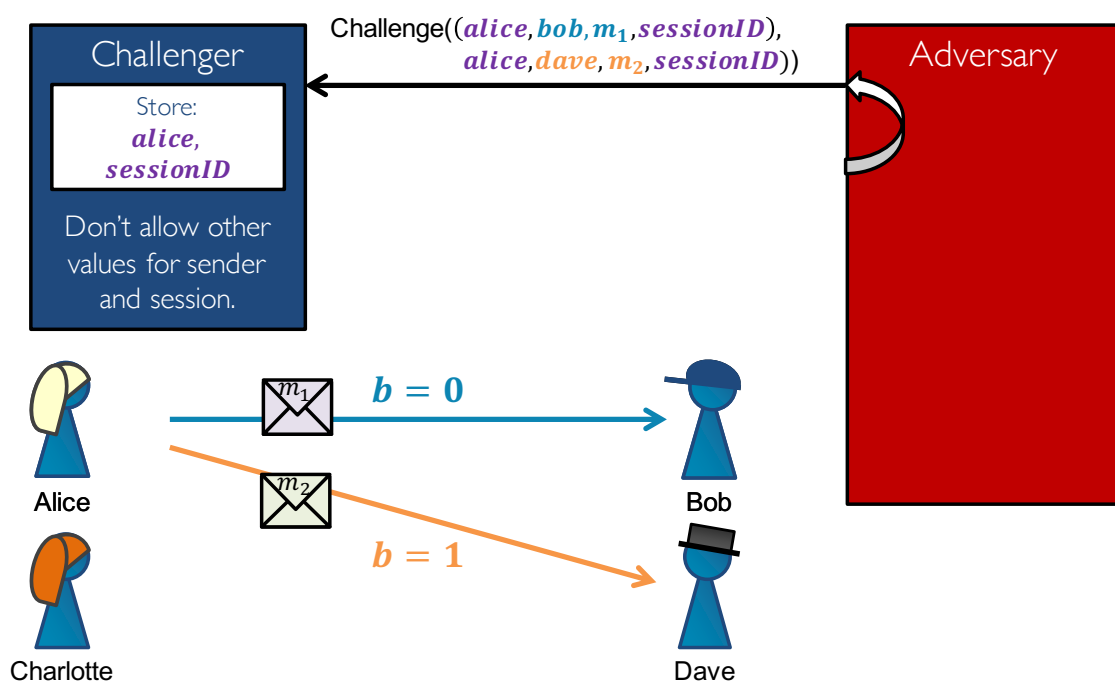
Single Message Sender Anonymity



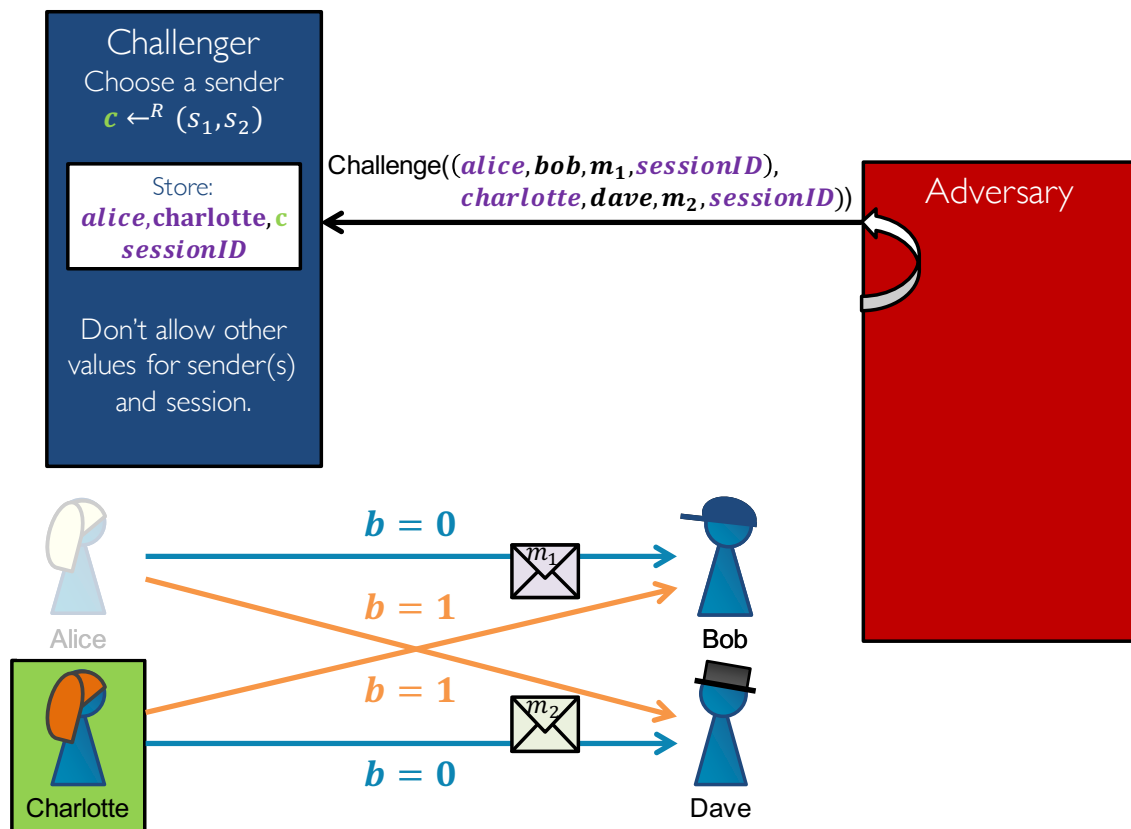
Session Sender Anonymity



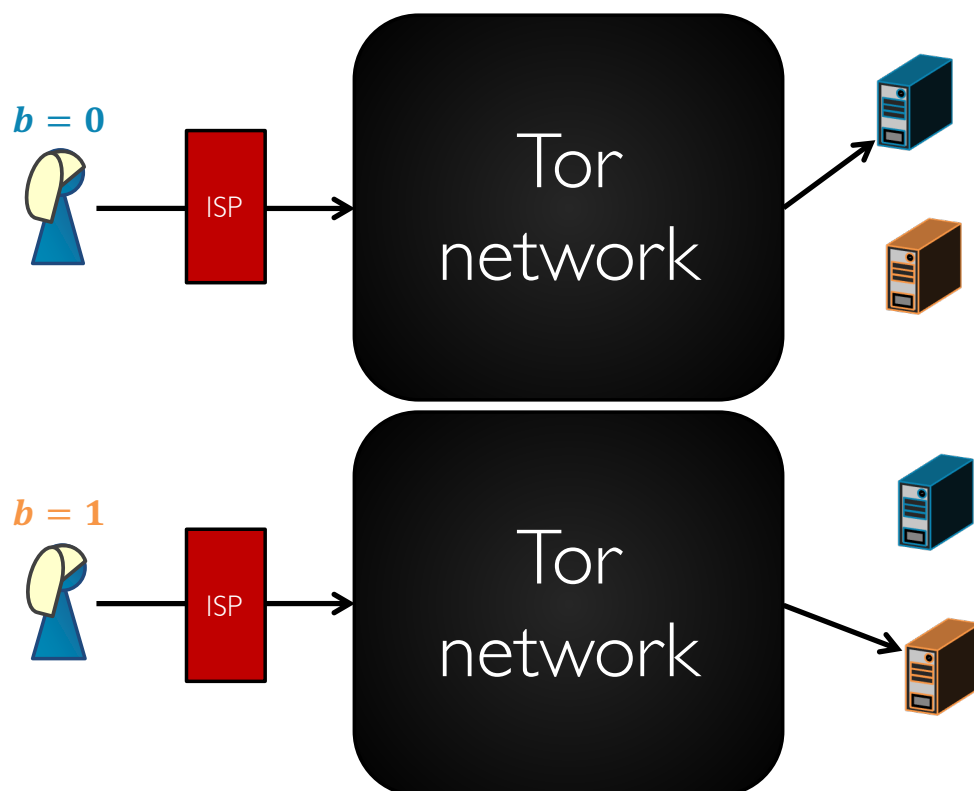
(Session) Recipient Anonymity



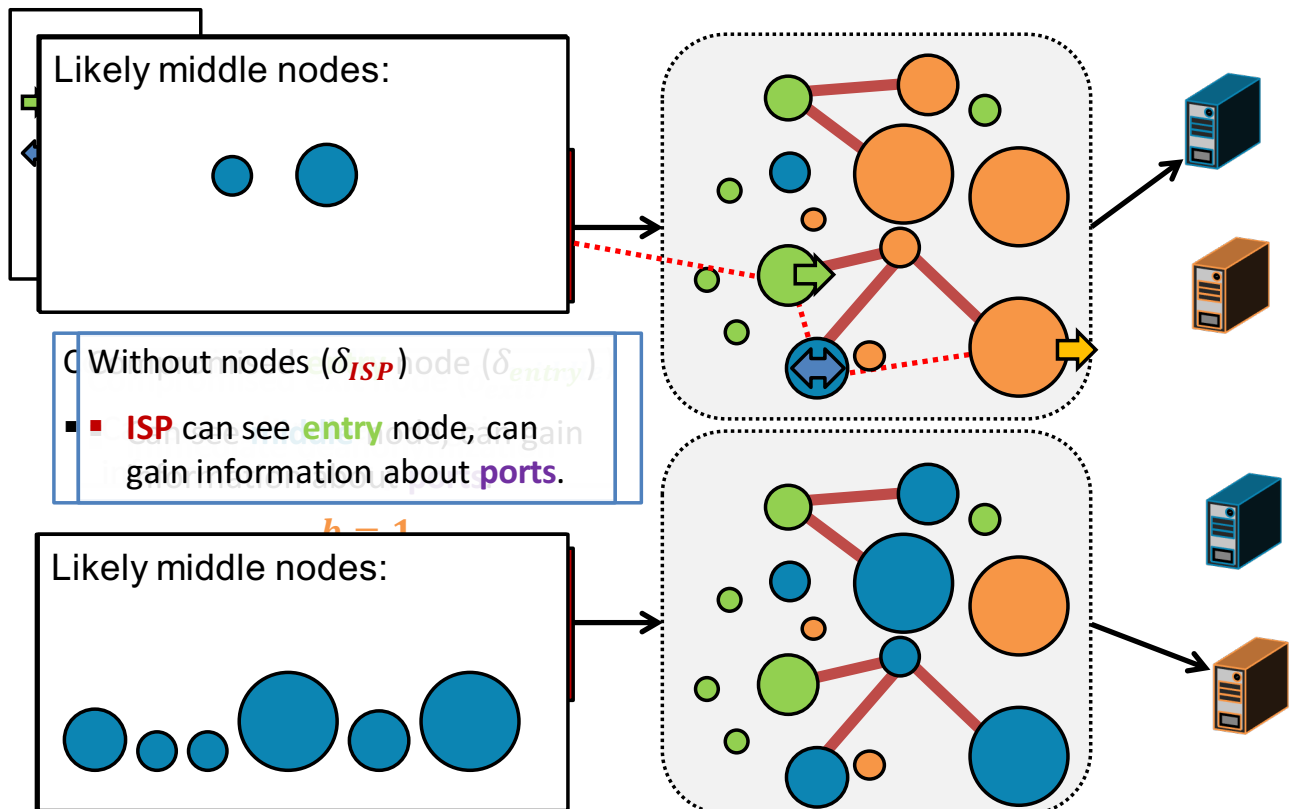
(Session) Relationship Anonymity



(Recipient) Anonymity for Tors Path Selection



(Recipient) Anonymity for Tors Path Selection



(Recipient) Anonymity for Tors Path Selection

$$\Pr[0 \leftarrow Adv | Chal^{b=0}] \leq \Pr[0 \leftarrow Adv | Chal^{b=1}] + \delta$$

Compromising **exit** node

- Immediate deanonymization

Compromising **middle** node

- Can see **exit** node, can gain information about **ports**.

Compromising **entry** node

- Can see **middle** node, can gain information about **ports**.

Without compromising nodes

- ISP can see **entry** node, can gain information about **ports**.

$$\delta \leq \delta_{exit} + \delta_{middle} + \delta_{entry} + \delta_{ISP}$$

(Recipient) Anonymity Formula

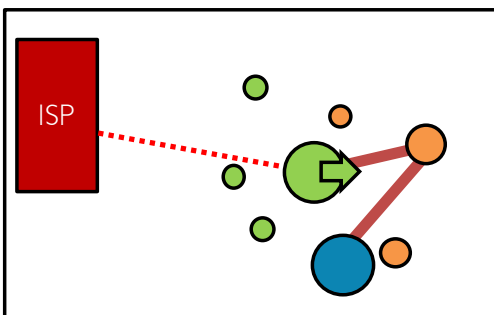
$$\Pr[0 \leftarrow Adv|Chal^{b=0}] \leq \Pr[0 \leftarrow Adv|Chal^{b=1}] + \delta$$

$\delta_{exit} + \delta_{middle} + \delta_{entry} + \delta_{ISP}$

$$\Pr[0 \leftarrow Adv|Chal^{b=0}] \leq \Pr[0 \leftarrow Adv|Chal^{b=1}] + x \cdot \Pr[0 \leftarrow Adv|Chal^{b=1}] + \delta'$$

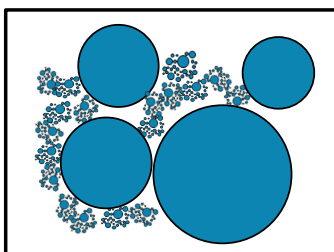
$$\Pr[0 \leftarrow Adv|Chal^{b=0}] \leq e^\epsilon \Pr[0 \leftarrow Adv|Chal^{b=1}] + \delta'$$

Significantly smaller.

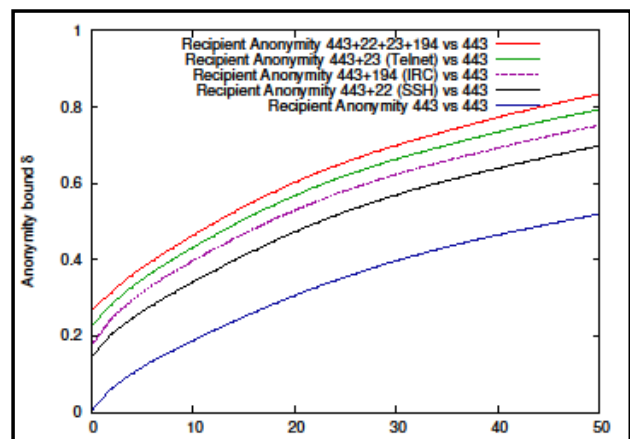


Lessons Learned

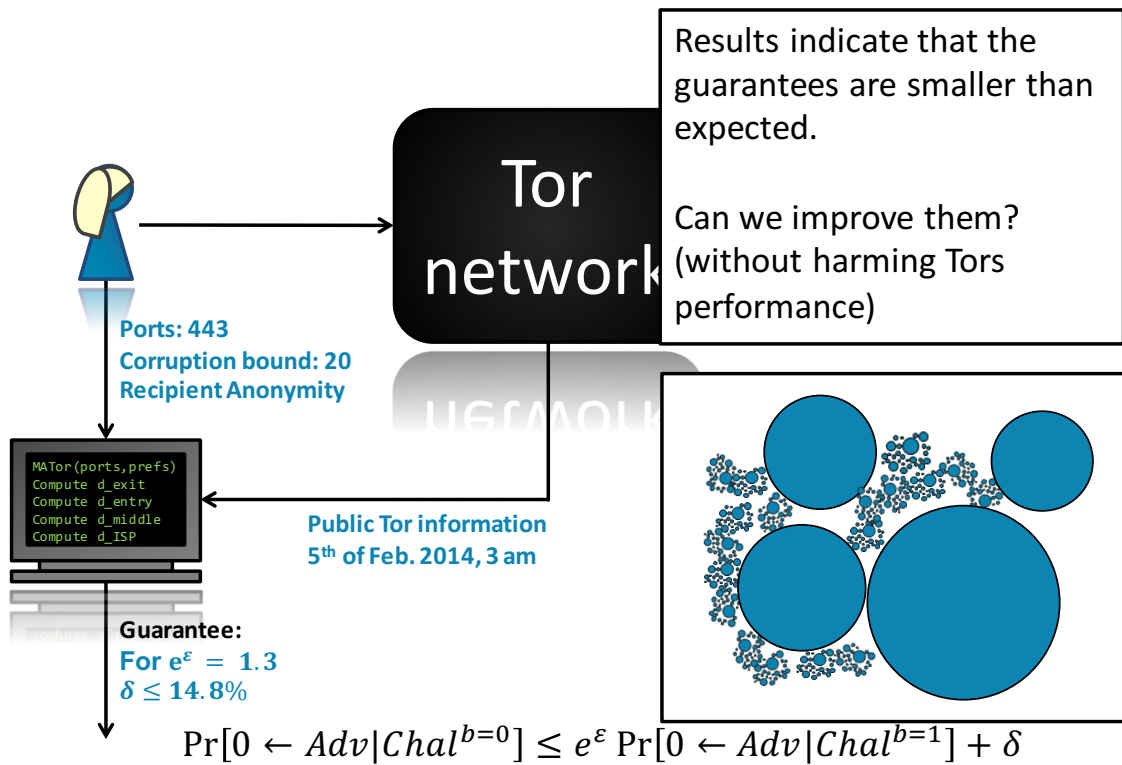
Ports play a significant role for recipient anonymity (and relationship anonymity)



Trust is not distributed evenly, but prioritizes very large nodes

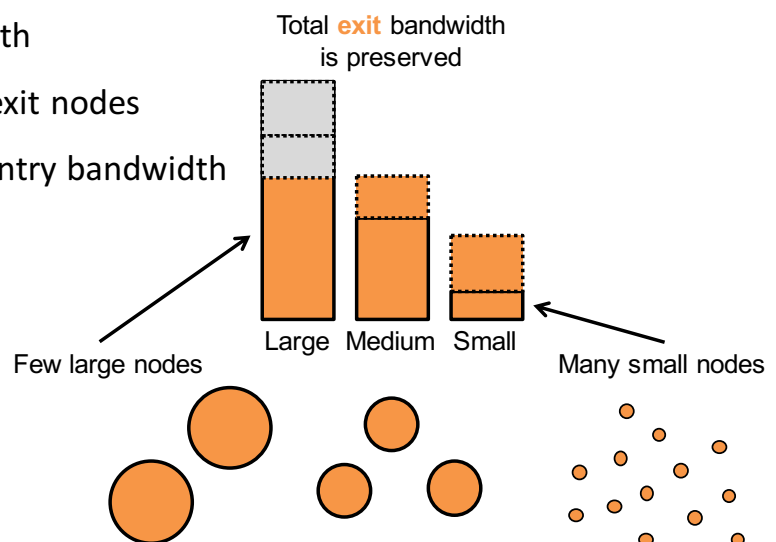


A live Monitor for Anonymity (of Tor's path selection)



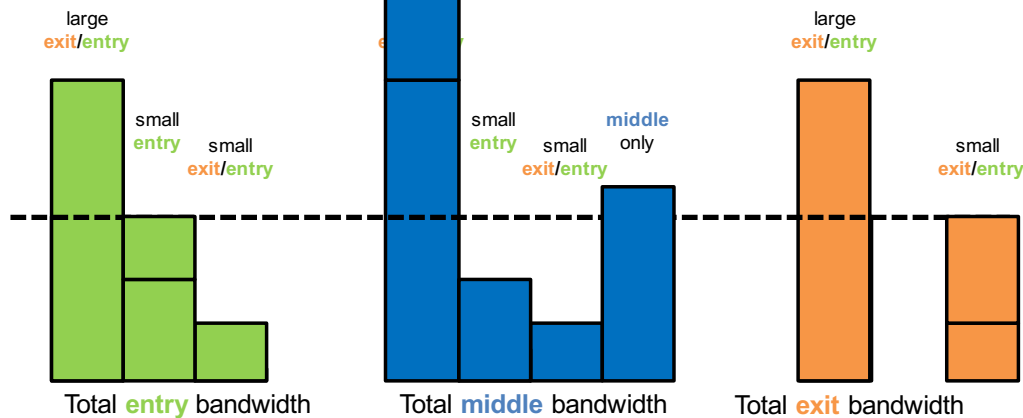
New Path Selection Algorithm: DistribuTor

- Goals:
 - As performant as Tor, while improving anonymity
 - Distribute the trust more evenly over as many nodes as possible
1. Compute new max. exit bandwidth
 2. Distribute the exit bandwidth
 3. Use smaller nodes only as exit nodes
 4. Then do the same for the entry bandwidth



New Path Selection Algorithm: DistribuTor

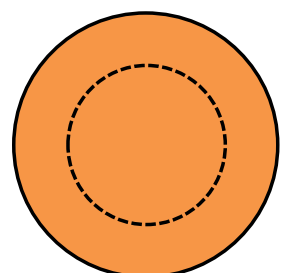
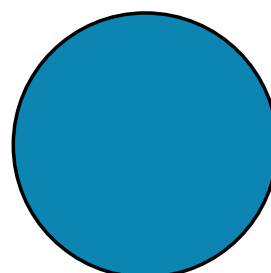
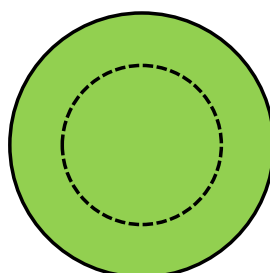
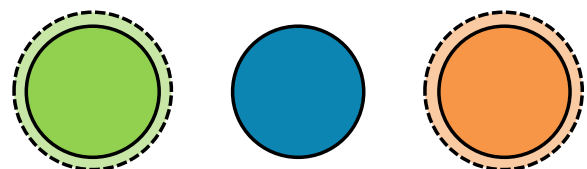
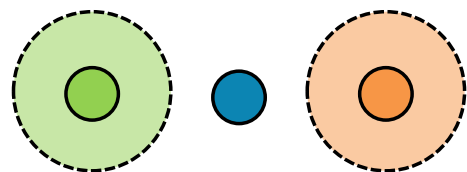
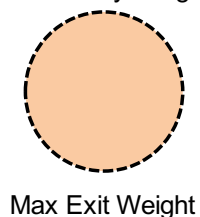
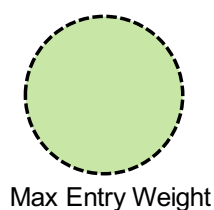
- Bandwidth modification per node
- First optimize for exit nodes, then optimize for entry nodes
- Use large nodes mainly as middle nodes



Contribution by large, medium, small and tiny nodes

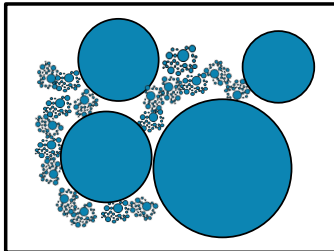
New Path Selection Algorithm: DistribuTor

- Don't reduce Tors overall performance
- Improve the anonymity bounds:
 - Redistribute the weights
 - Use small exit nodes only as exit nodes
 - Restrict the usage of large nodes



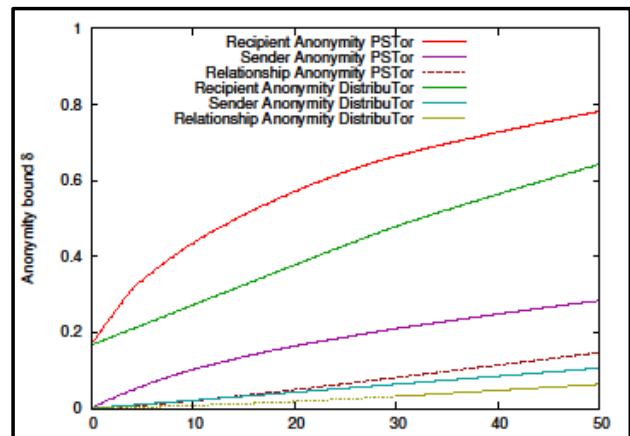
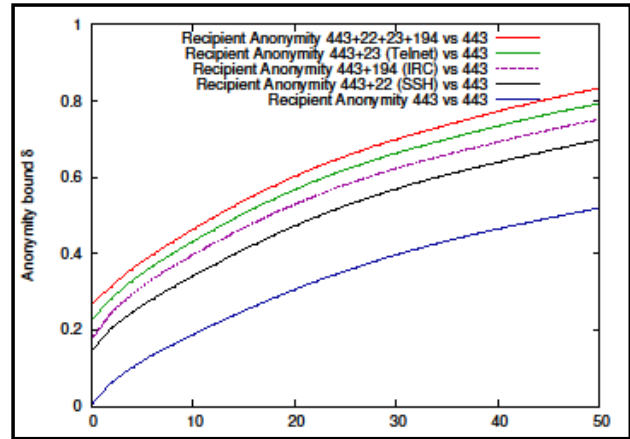
Lessons Learned

Ports play a significant role for recipient anonymity (and relationship anonymity)



Trust is not distributed evenly, but prioritizes very large nodes

DistribuTor achieves better guarantees without reducing the performance

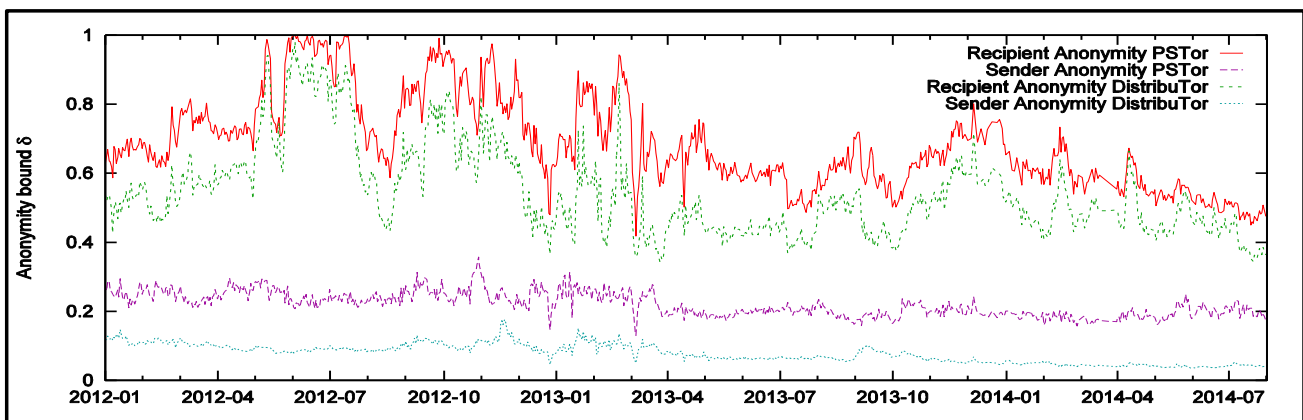
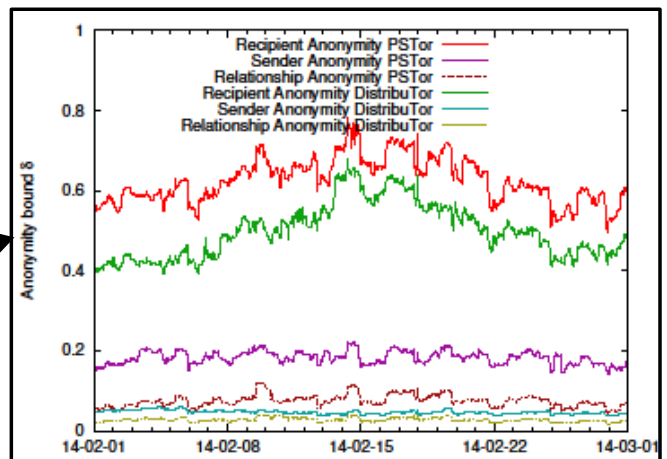


Lessons Learned

The guarantees change over **time**, even on an hourly basis

Guarantee during **February 2014**

Guarantee from **January 2012 until July 2014**



Next Steps

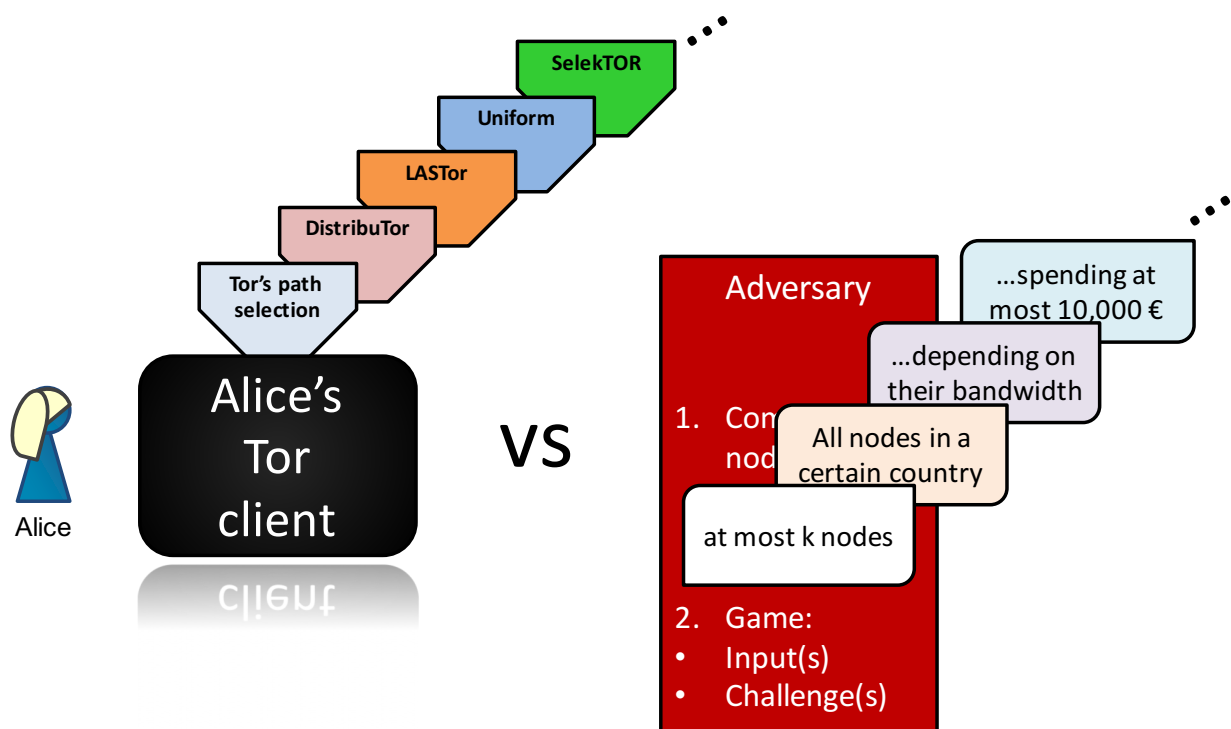
Performance

Preparation (PSA): 3.39s
Sender anonymity: 0.73s
Recipient anonymity: 6.07s
Relationship anonymity: 9.10s

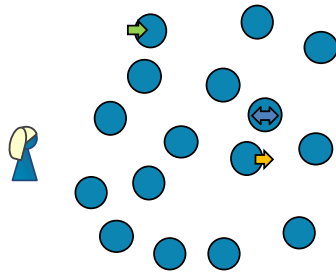
Parameters of our Analysis:

- (here): Ports, Exit/Guard Flags, Tors Path Selection, DistribuTor, Recipient Anonymity against k-of-N adversaries
- (in the Paper): Sender Anonymity, Relationship Anonymity against k-of-N adversaries
- (current work): More path selection algorithms,
 - How dangerous is the conversion phase?
 - What if an adversary compromises based on BW, not k of N?
 - What if all nodes within a country are compromised?
 - ...

Generalization of the Anonymity Analysis



Generalized Notion of Path Selection Algorithms



- Model not only Tors current path selection algorithm, but also alternatives.
 - LASTor
 - Uniform path selection
 - Modifications of Tors path selection algorithm

Path Selection:

(remaining assumptions)

1. Choose a possible **exit** node that allows Alice's **ports**.
2. Choose a possible **entry** node, depending on the **exit** node.
3. Choose a possible **middle** node, depending on **exit** and **entry** node.

Generalizing Adversaries

Adversary Classes:

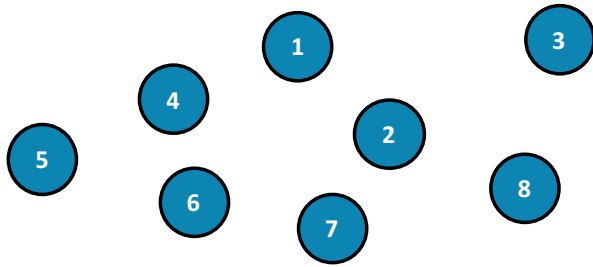
- Filter information from the network learned by the adversary
- Restrict actions that can be performed by the adversary



- Different adversary classes for different real adversaries:
 - Budget-adversary
 - Infrastructure adversary

Advanced Compromisation Strategies

- Adversary strategy described by cost function f and budget B



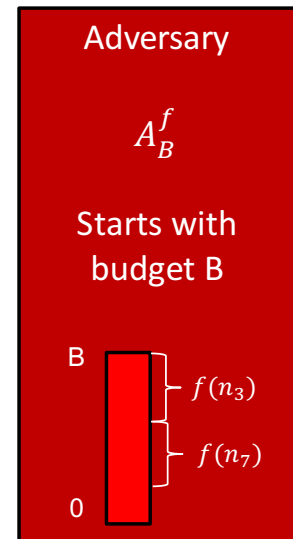
- Examples:

- Adversary compromises all nodes in the Netherlands:

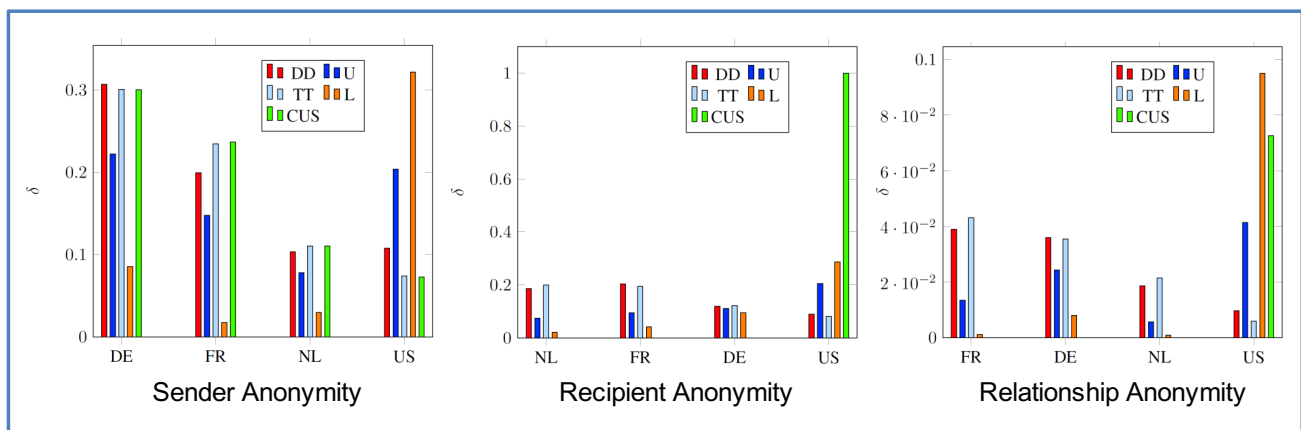
$$f_{\text{country-NL}}(n) = \begin{cases} 0 & \text{if } n \in \text{NL} \\ \infty & \text{otherwise} \end{cases}$$

- Adversary compromises based on bandwidth:

$$f_{\text{bandwidth}}(n) = n.\text{bandwidth}$$



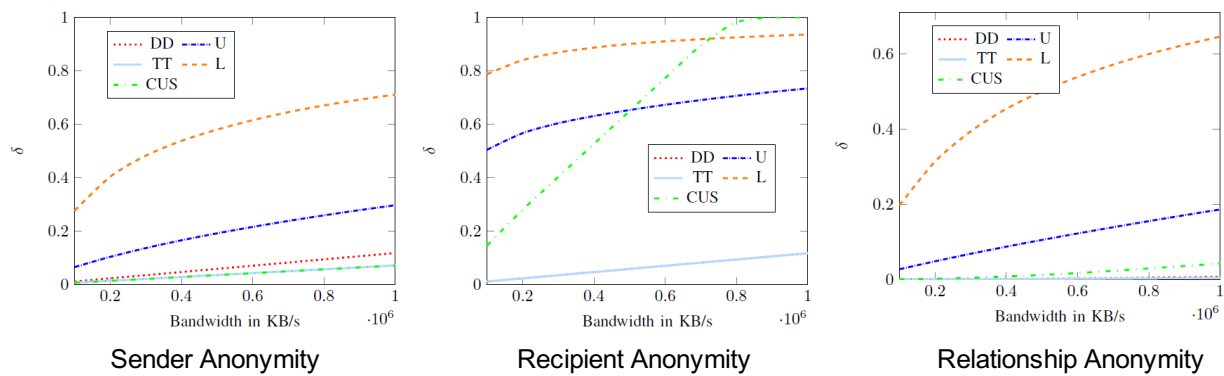
Compromised Nodes in Countries



- Path selection algorithms:

- Tor's path selection (TT)
- DistribuTor (DD)
- LASTor (L) – Chooses nodes depending on geographical distances
- Uniform (U) – Chooses every node uniformly at random
- Selektor [only US exit] (CUS) – Only picks exit nodes from the US

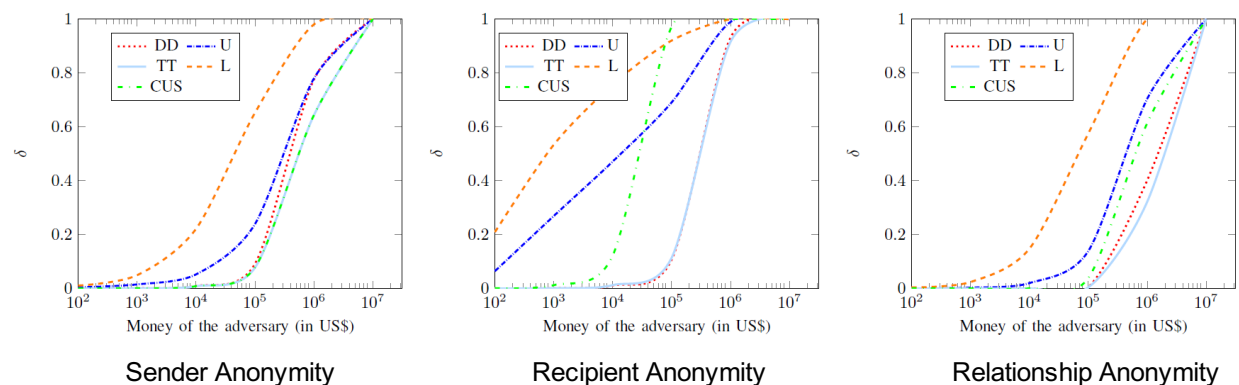
Compromised Bandwidth



Path selection algorithms:

- Tor's path selection (**TT**)
- DistribuTor (**DD**)
- LASTor (**L**) – Chooses nodes depending on geographical distances
- Uniform (**U**) – Chooses every node uniformly at random
- Selektor [only US exit] (**CUS**) – Only picks exit nodes from the US

Compromisation by Costs

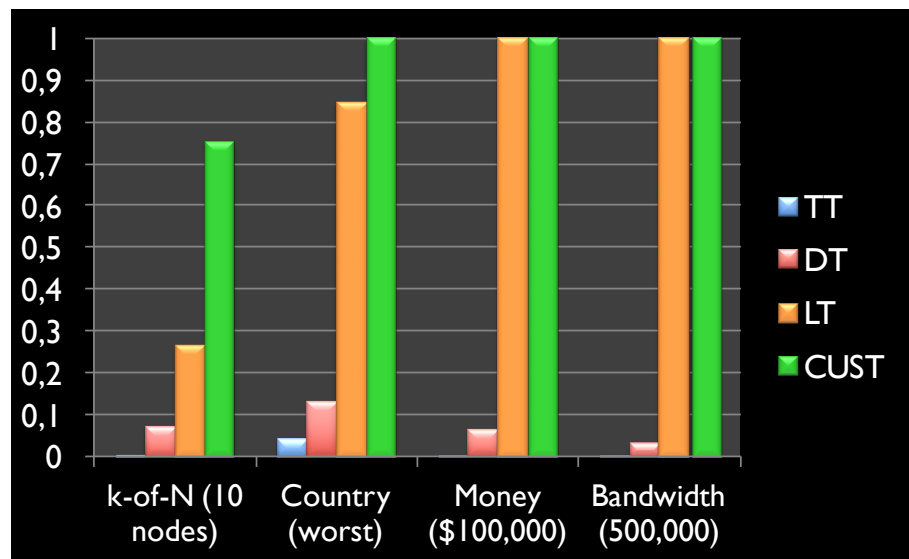


Path selection algorithms:

- Tor's path selection (**TT**)
- DistribuTor (**DD**)
- LASTor (**L**) – Chooses nodes depending on geographical distances
- Uniform (**U**) – Chooses every node uniformly at random
- Selektor [only US exit] (**CUS**) – Only picks exit nodes from the US

Conversion Phase:

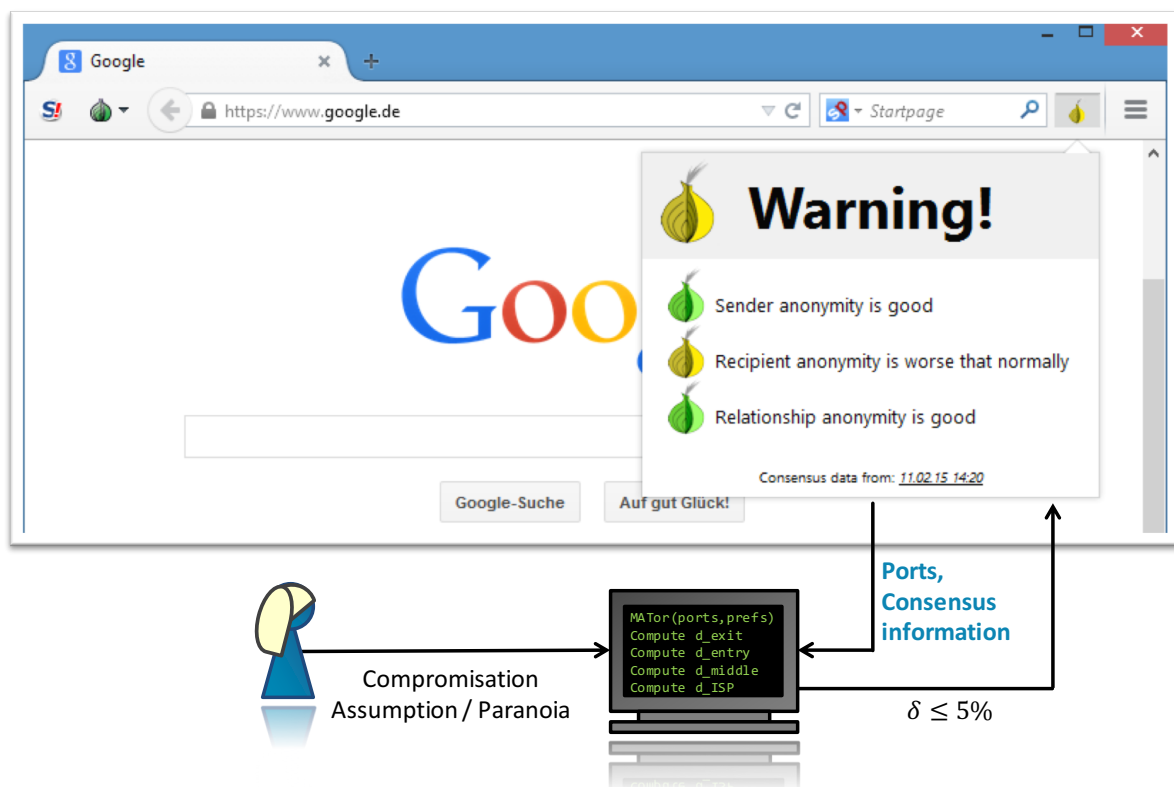
Using a Non-Standard Path Selection Algorithm



Scenarios:

- Tor's Path selection (**TT**) (for comparison)
- Using Distributor when everyone else uses Tor's path selection (**DT**)
- Using LASTor when everyone else uses Tor's path selection (**LT**)
- Using Selektor [only US exit] when everyone else uses Tor's path selection (**CUS**)

Ongoing Work: Integration into the Tor Browser



Anonymity Impact of Malicious Infrastructure

1. Review: Internet Infrastructure
2. Assessing Anonymity with malicious infrastructure
3. Computing Anonymity Guarantees under malicious infrastructure
4. Reconstructing Internet routing relevant for Tor
5. Choosing a representative, connected component of Tor
6. Experimental Evaluation

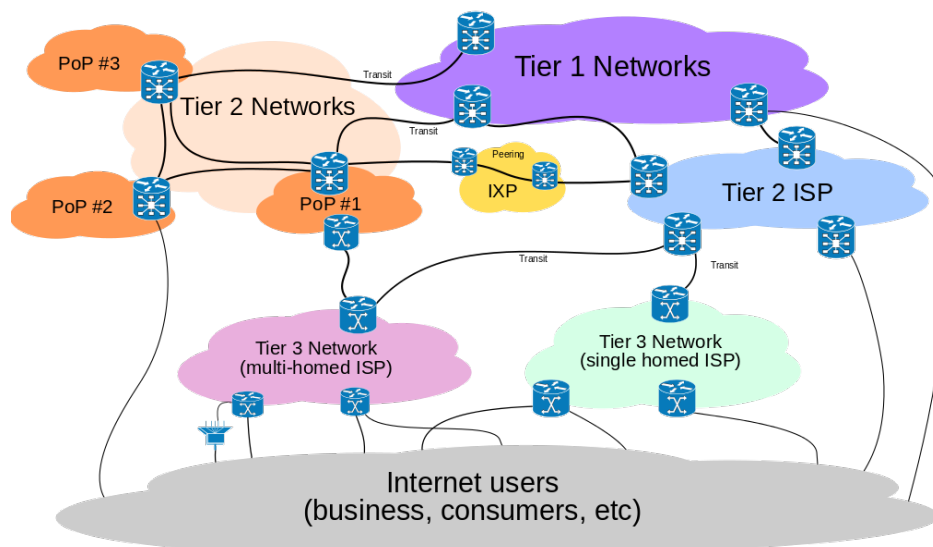
Tor is an Overlay Network

- Recall: Tor is an overlay network over the Internet
- Understanding the travel path of data requires understanding the topology of the Internet, i.e., how data is routed in the Internet



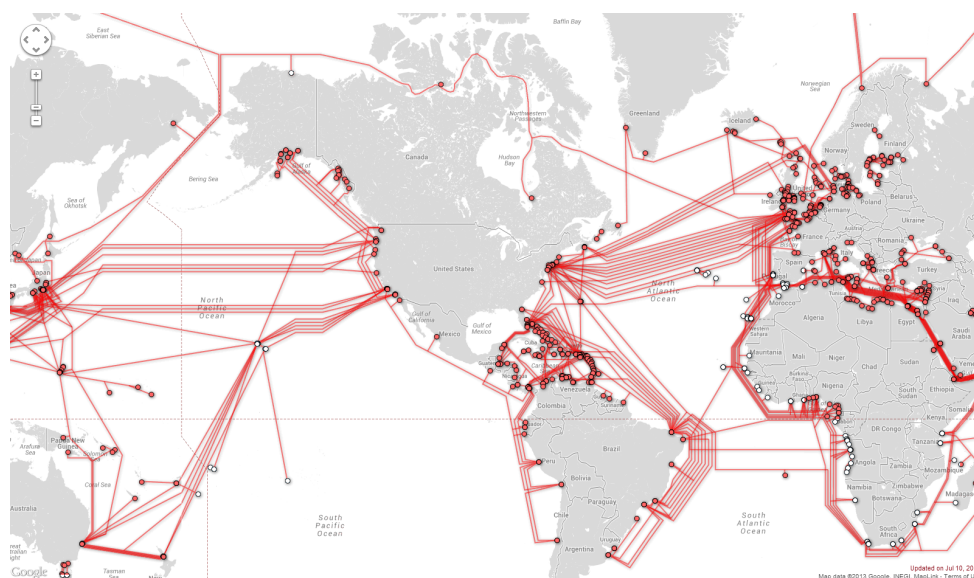
1. Review: Internet Infrastructure

- Lower-tier autonomous systems (ASes) connect to higher-tier ASes (networks in the graphic)
- Recently, Internet Exchange Points (IXPs) also directly connect lower-tier provider
- Points of presence (PoP) are the locations where cables are physically connected
- Routers at the border of Networks, IXPs, and PoPs



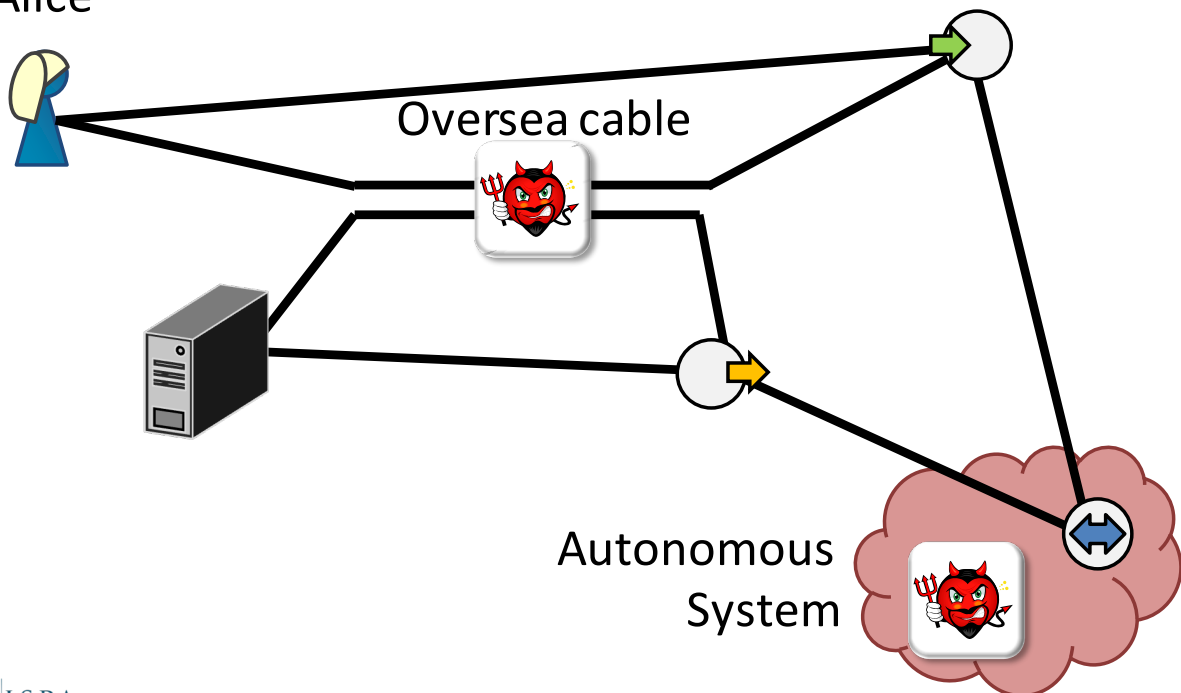
Submarine cables

- Submarine cables connect continents through the oceans
- Submarines have several landing points where a cable is connected to landline networks
- Some cables are government-supported (TAT-14) but most cables are owned by multi-national corporations (e.g., Level3, Verizon, DTAG, NTT, BT)

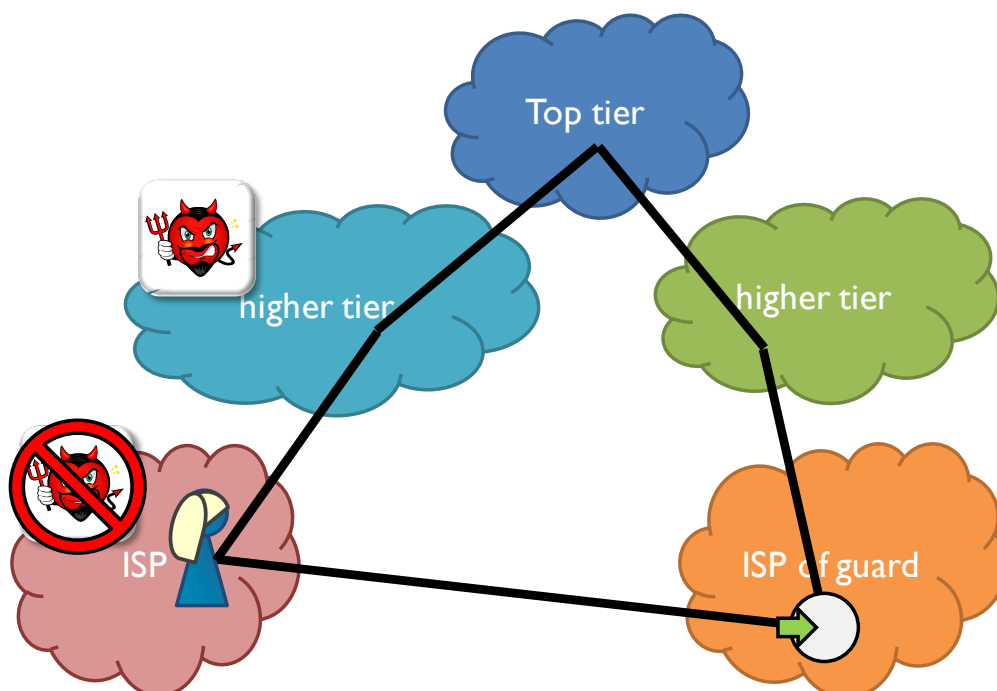


Network infrastructure level attacks (work in process)

Alice



Network infrastructure level attacks Worst case anonymity is impossible/meaningless

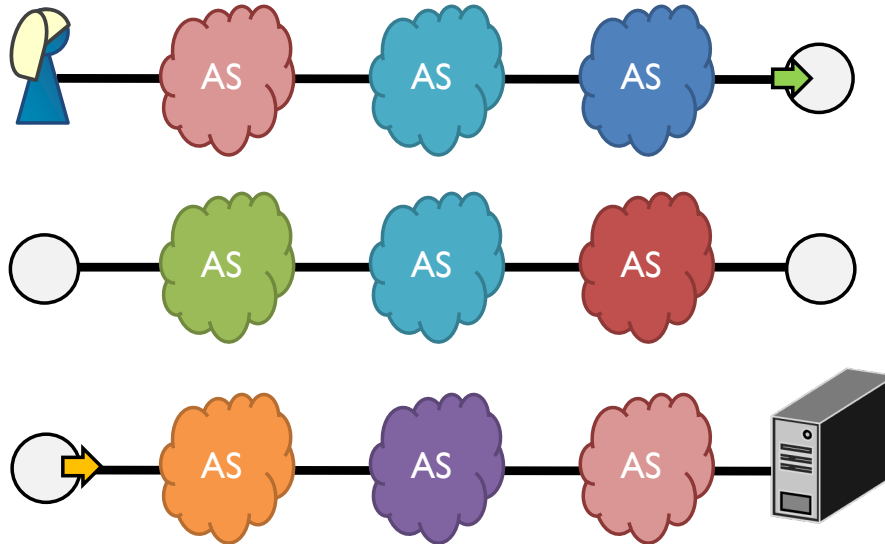


A worst-case adversary would always compromise Alice's ISP.

If we disallow that, it will compromise the next higher peers of Alice's ISP.

Network infrastructure level attacks (work in process)

Alice



Find out the autonomous systems through which traffic is routed. Either via traceroute or via BGP messages.

2. Assessing Anonymity with Malicious Infrastructure

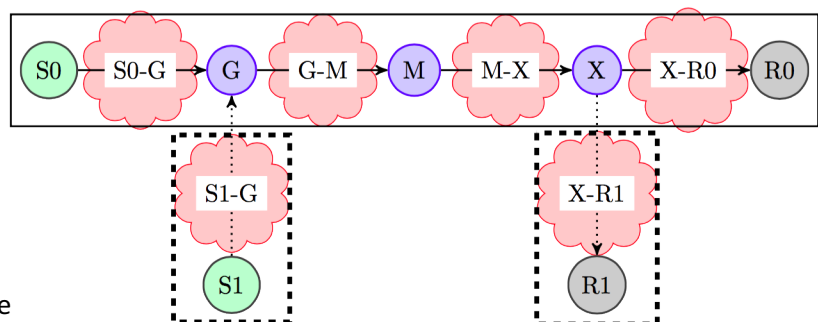
- Possible Observations points of malicious infrastructure
- Recall that relationship anonymity involves four scenarios

⇒ i) (S0,G,M,X,R0); ii) (S1,G,M,X,R1);
iii) (S0,G,M,X,R1); iv) (S1,G,M,X,R0)

- Traffic correlation attacks enable the combination of observations:

⇒ e.g., S0-G plus M-X for the same circuit can be combined to $S0-G \perp M-X \perp$
(\perp denotes that in the combined observation for this point, there is no observation)

- Given a topology of the internet
- Given a malicious infrastructure (i.e., which parts of the topology is malicious)
- We can for each Tor circuit (G,M,X) compute the observable points



- Senders S0, S1
- Entry guard G
- Middle node M
- Exit node X
- Recipients R0, R1

3. Computing Anonymity Guarantees

```

COMPUTEANONYMITY( $\mathcal{MI}$ ,  $S0$ ,  $S1$ ,  $R0$ ,  $R1$ )
  for all  $z \in \{S0, S1\} \times \{R0, R1\} \times Obs(S0, S1, R0, R1)$  do
    store[ $z$ ] := 0
  store := OBSERVATIONPHASE(store,  $\mathcal{MI}$ ,  $S0$ ,  $S1$ ,  $R0$ ,  $R1$ )
  ( $\delta_{SA}$ ,  $\delta_{RA}$ ,  $\delta_{REL}$ ) := DEDUCTIONPHASE(store,  $S0$ ,  $S1$ ,  $R0$ ,  $R1$ )
  return ( $\delta_{SA}$ ,  $\delta_{RA}$ ,  $\delta_{REL}$ )

```

- Compute the probability of each (combined) observation in the four scenarios
- Observation Phase:
 - compute the probability of an observation in each of the four scenarios
 - store contains the probabilities for each these observations
- Deduction Phase:
 - Compute the adversary's advantage out of these probabilities

Computing Anonymity Guarantees: Observation Phase

```

OBSERVATIONPHASE(store,  $\mathcal{MI}$ ,  $S0$ ,  $S1$ ,  $R0$ ,  $R1$ )
  for all  $(s, r) \in \{S0, S1\} \times \{R0, R1\}$  do
     $P_{s,r} := \text{new TORPS}(s, r)$ 
  for all  $(s, r) \in \{S0, S1\} \times \{R0, R1\}, (n_G, n_M, n_X) \in \mathcal{N}^3$  do
    store[ $s, r, OBS(\mathcal{MI}, s, n_G, n_M, n_X, r)$ ] +=  $P_{s,r}(n_G, n_M, n_X)$ 
  return store

OBS( $\mathcal{MI}$ ,  $s$ ,  $n_G$ ,  $n_M$ ,  $n_X$ ,  $r$ )
  Initialize  $i := \perp$  for  $i \in \{o_s, o_G, o_M, o_X, o_r\}$ 
  if  $(s, n_G) \in \mathcal{MI}$  then  $o_s := s; o_G := n_G$ 
  if  $(n_G, n_M) \in \mathcal{MI}$  then  $o_G := n_G; o_M := n_M$ 
  if  $(n_M, n_X) \in \mathcal{MI}$  then  $o_M := n_M; o_X := n_X$ 
  if  $(n_X, r) \in \mathcal{MI}$  then  $o_X := n_X; o_r := r$ 
  return ( $o_s, o_G, o_M, o_X, o_r$ )

```

- Represent path selection strategy as probability distribution $P_{s,r}$:
 - each node combination has some probability that it is chosen

Computing Anonymity Guarantees: Deduction Phase

```
DEDUCTIONPHASE(store, S0, S1, R0, R1)
   $\delta_{SA}, \delta_{RA}, \delta_{REL} := 0$ 
  for all  $o \in Obs(S0, S1, R0, R1)$  do
    ADDDIFF( $\delta_{SA}, store[S0, R0, o], store[S1, R0, o]$ )
    ADDDIFF( $\delta_{RA}, store[S0, R0, o], store[S0, R1, o]$ )
     $r_1 := (store[S0, R0, o] + store[S1, R1, o])/2$ 
     $r_2 := (store[S0, R1, o] + store[S1, R0, o])/2$ 
    ADDDIFF( $\delta_{REL}, r_1, r_2$ )
  return ( $\delta_{SA}, \delta_{RA}, \delta_{REL}$ )

ADDDIFF(Z, X, Y)
  if  $X > Y$  then
     $Z += X - Y$ 
```

- Add up the differences of the probabilities for the four scenarios
- We only consider one direction (e.g., S0,R0 minus S1, R0) since the opposite direction (e.g., S1, R0 minus S0, R0) can be proven to be the same

Representation in AnoA

- Represent a malicious infrastructure as a set MI of
 - pairs of Tor nodes,
 - pairs of senders and Tor nodes, and
 - pairs of Tor nodes and recipients
- Malicious Infrastructure adversaries can be represented as a specific family of adversary classes, parametric in MI:
 - Upon receiving an observation from the challenger (which an omniscient adversary would be able to make):
 - If the involved nodes of that observation are in MI, forward the observation to the adversary
 - Forward all other messages in both directions

Correctness Proof

- We can show:
 - The algorithm Compute Anonymity precisely computes the advantage of the adversary.
- Main insight: the combined observation points are elementary probability events for the success probability of the adversary (in each of the settings, respectively)

How to get the real Internet topology?

Can we find out how data is routed through the Internet?

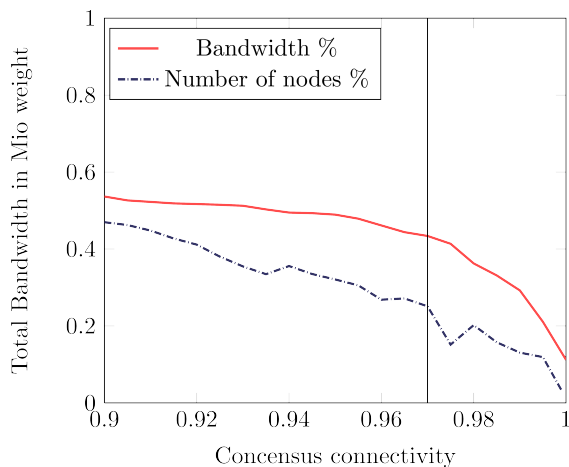
4. Reconstructing Internet Routing Relevant for Tor

- Little public data
 - For a few ASes it is known with whom they peer
 - Even less ASes publish coarse-grained routing paths (BGP paths)
 - Hard to predict routing policies
 - Geographic proximity of the subnet is taken into account
 - Contracts between the ASes are taken into account
 - Routing based on subnets
- ⇒ We use the established service iPlane [1] for gathering routing information
- Conducts measurements
 - extrapolates routing information from these measurement and public BGP data

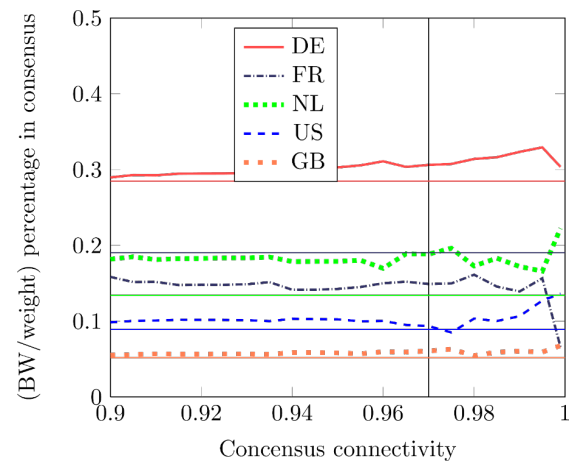
Working with Incomplete Data

- iPlane is inherently incomplete
- We only got a fraction of the Internet routing information
- We also only got a fraction of the routing information relevant for Tor
- Approach:
 - Find a subset of Tor nodes that is
 - covered by our routing information
 - representative in terms of bandwidth, country distribution, etc.
- We found a snapshot of the Tor Network consisting of 1650 nodes with 97% completeness of routing information

5. Choosing a representative, connected component of Tor



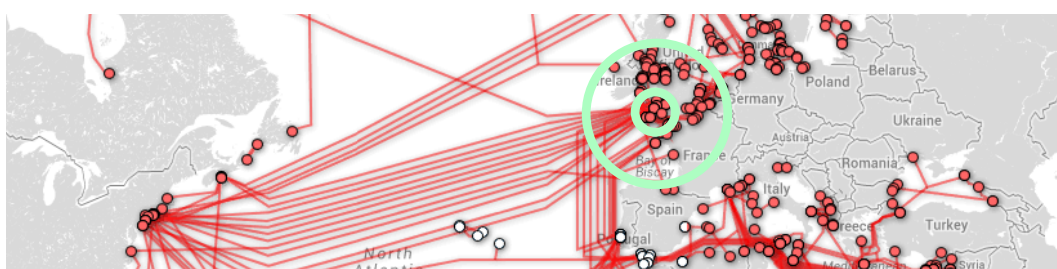
Total Bandwidth covered by the subset



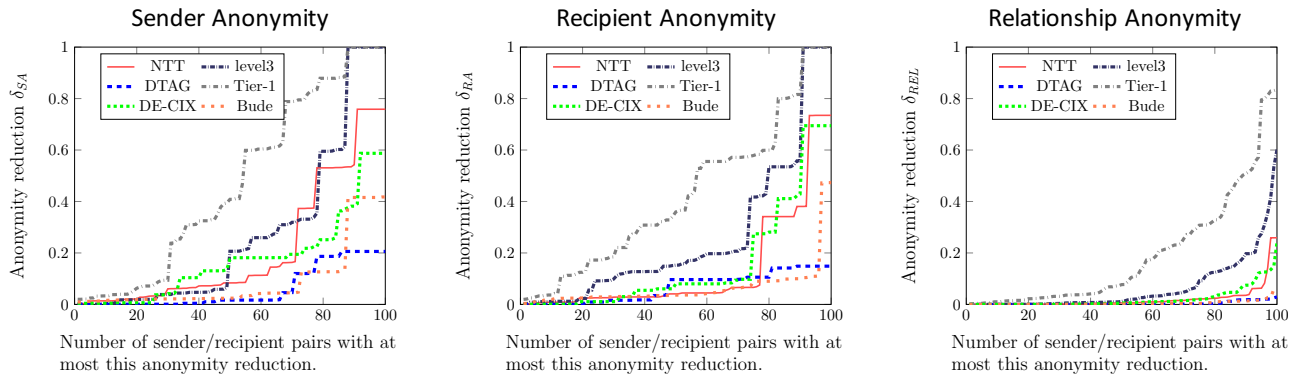
Bandwidth distribution per country

6. Experimental Evaluation

- We considered six adversaries
 - NTT: all ASes that belong to Nippon Telegraph and Telephone (NTT)
 - Level 3: all ASes that belong to Level 3
 - DTAG: all ASes that belong to Deutsche Telekom AG (DTAG)
 - Tier 1: The combination of NTT, Level3, and DTAG
 - DE-CIX: DE-CIX is the world's largest IXP and located in Frankfurt, many european subnets are connected to the DE-CIX
 - Bude landing point: The Bude landing point is located in West-England and provides a landing point for many transcontinental submarine cables.

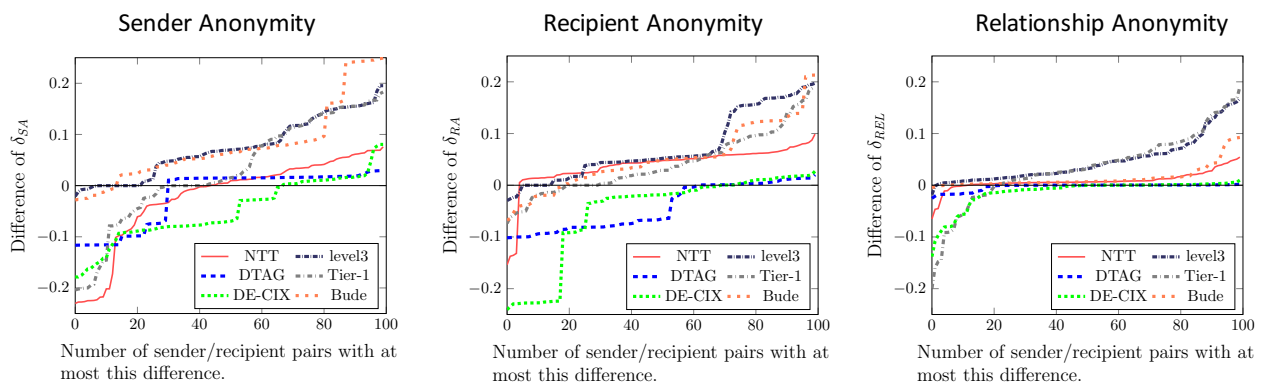


Tor's Path Selection



- We chose 20 users/recipients and sampled 100 combinations S_0, S_1, R_0, R_1 from these
- The graphs show how much advantage the adversary has at most for how many of the combinations
- These graphs depicts the results for the scenarios where both senders request the port 443 (HTTPS)

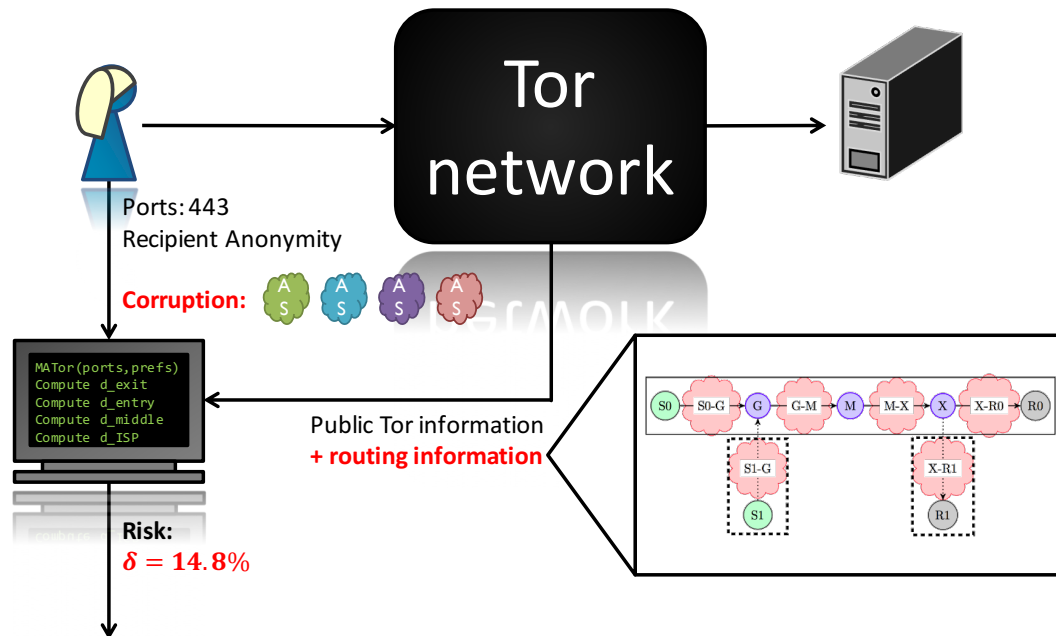
Difference LASTor



- The graphs show the **difference** if both senders use the path selection algorithm LASTor
- We chose 20 users/recipients and sampled 100 combinations S_0, S_1, R_0, R_1 from these
- The graphs show how much advantage the adversary has at most for how many of the combinations
- These graphs depicts the results for the scenarios where both senders request the port 443 (HTTPS)

Summary: MATor for Malicious Infrastructure

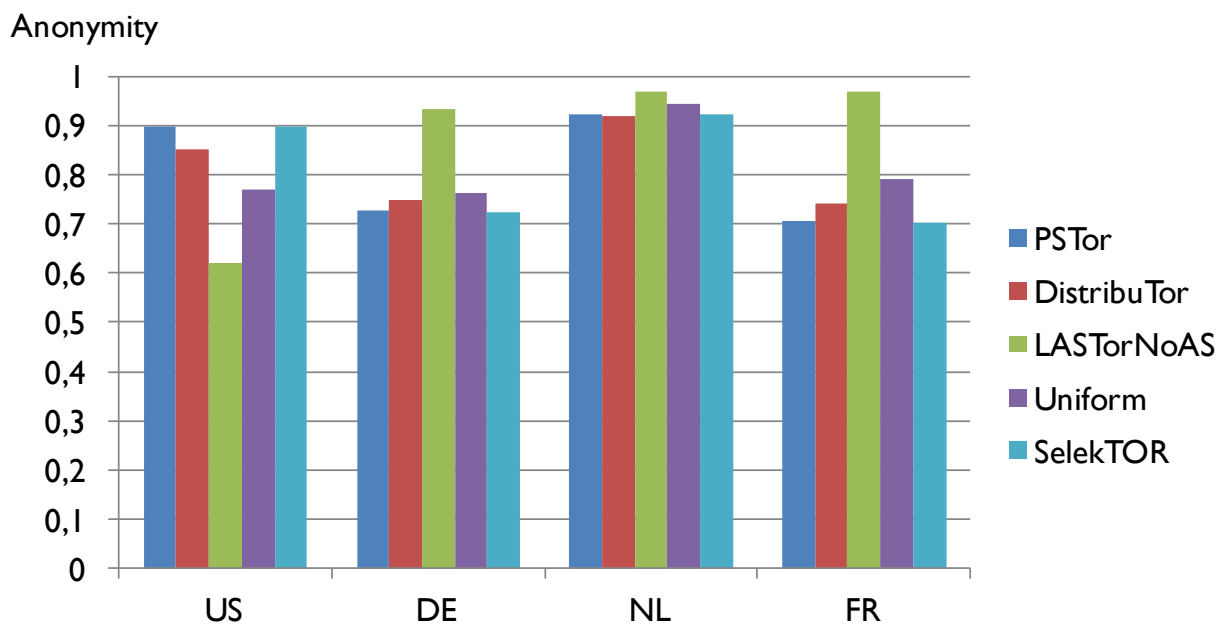
- Malicious Infrastructure:
 - Tier 1 providers (NTT, Level 3, DTAG, combination), Submarine Cable Landing Point (West-England), DE-CIX (world's largest IXP)



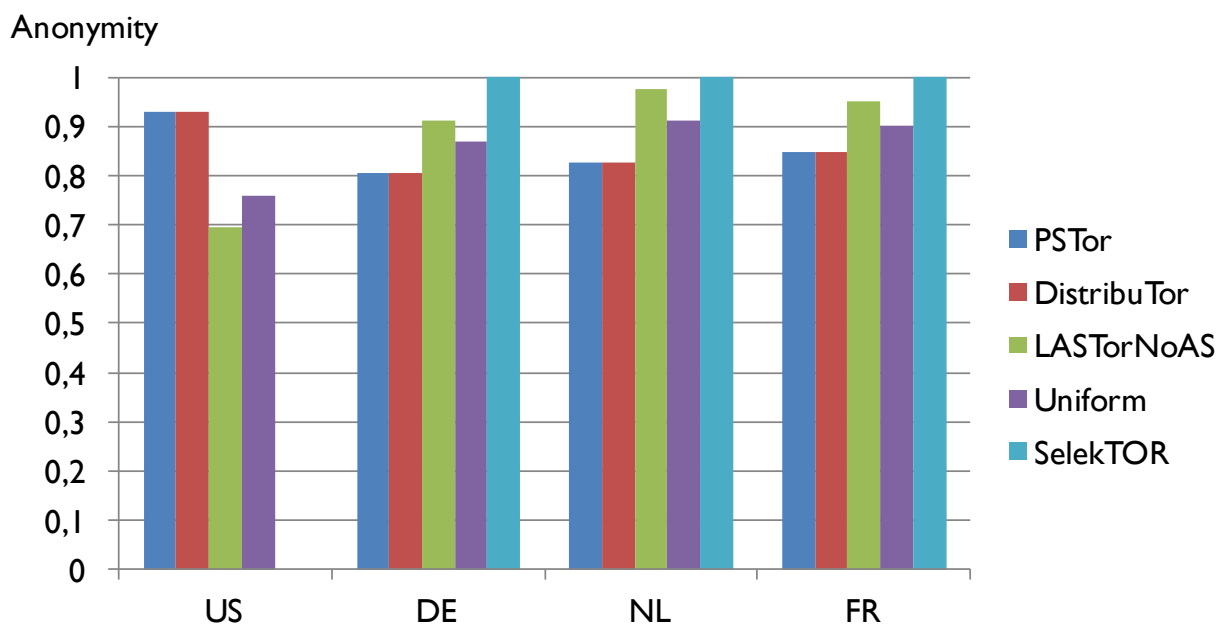
INFO FOR MICHAEL

- The following slides show the graphs in a more understandable manner.
- They do not show the advantage (δ) of the adversary, but the anonymity ($1 - \delta$)

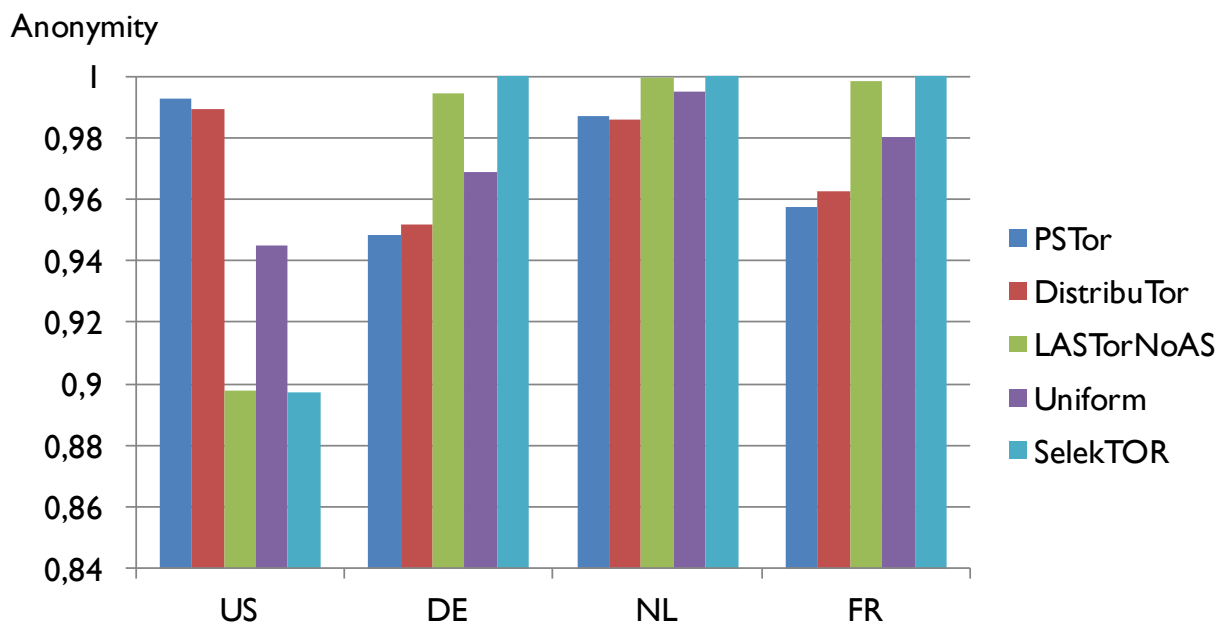
Compromised Nodes in Countries (Sender Anonymity)



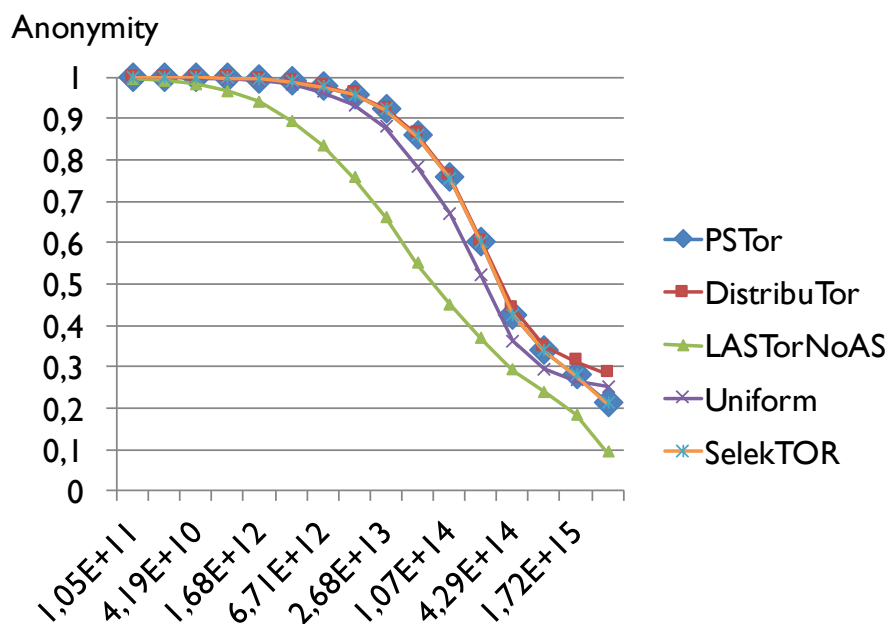
Compromised Nodes in Countries (Recipient Anonymity)



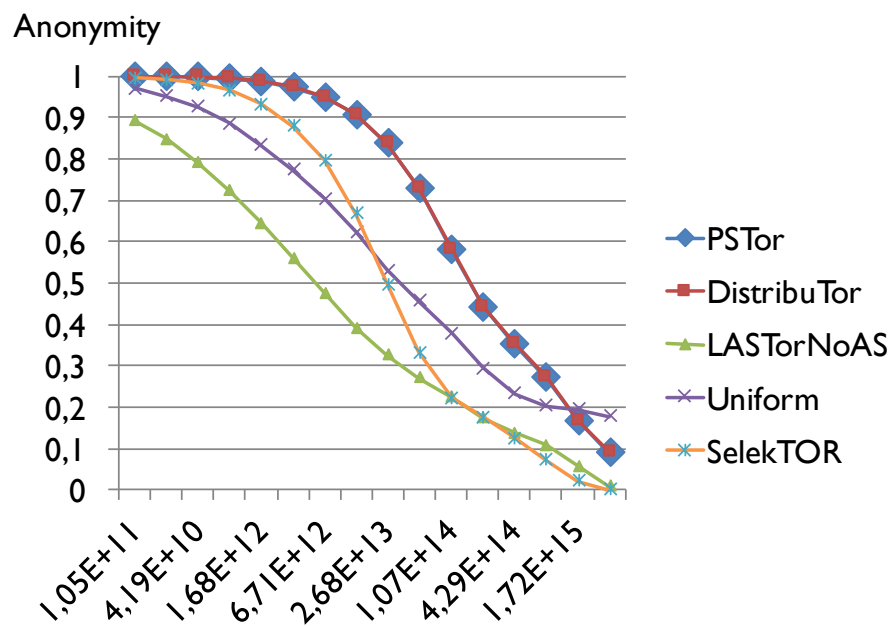
Compromised Nodes in Countries (Relationship Anonymity)



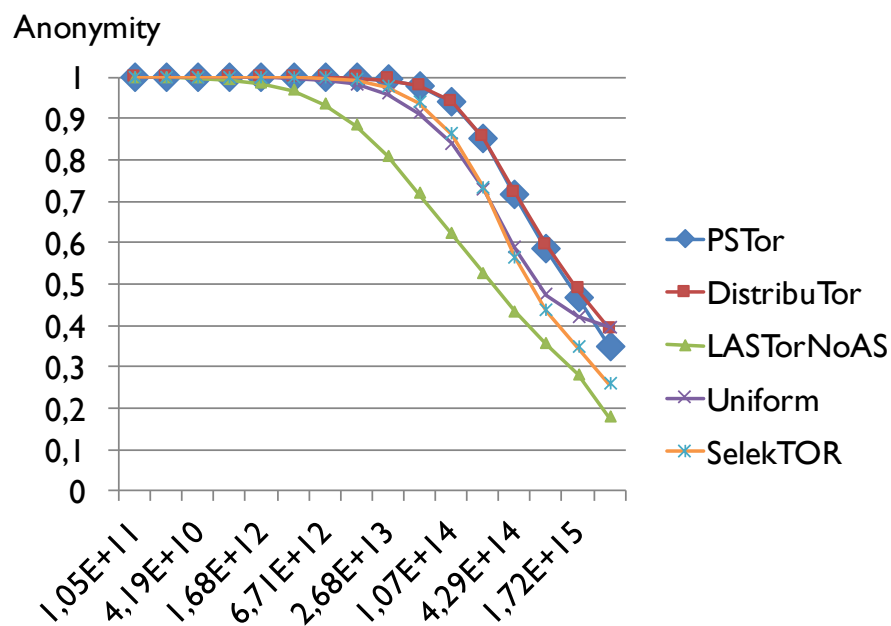
Bandwidth Adversary (Sender Anonymity)



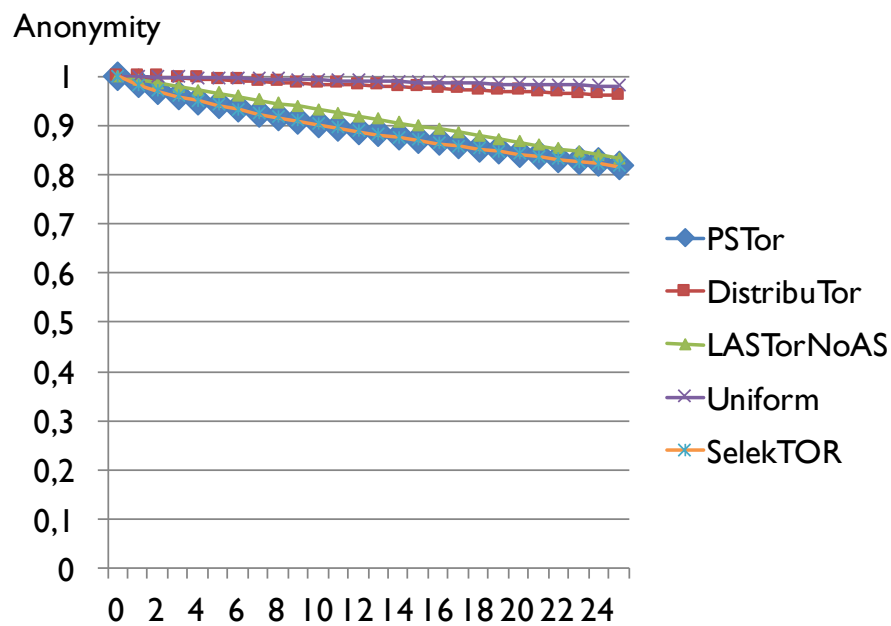
Bandwidth Adversary (Recipient Anonymity)



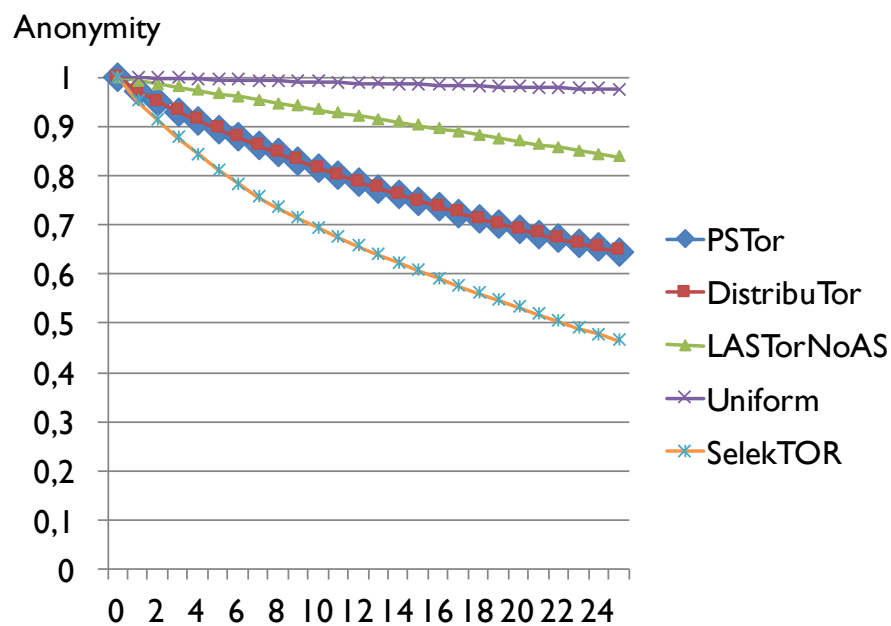
Bandwidth Adversary (Relationship Anonymity)



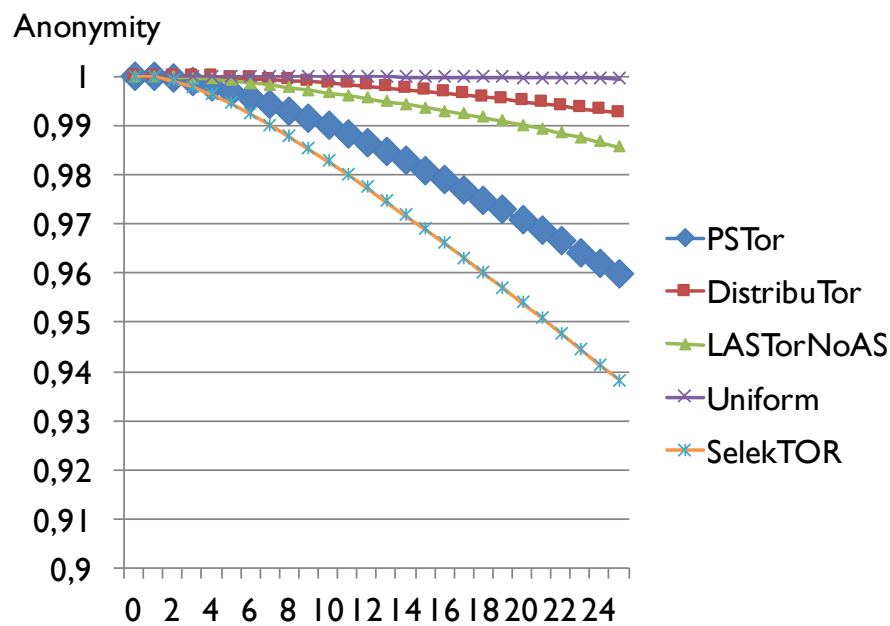
Byzantine Adversary (Sender Anonymity)



Byzantine Adversary (Recipient Anonymity)



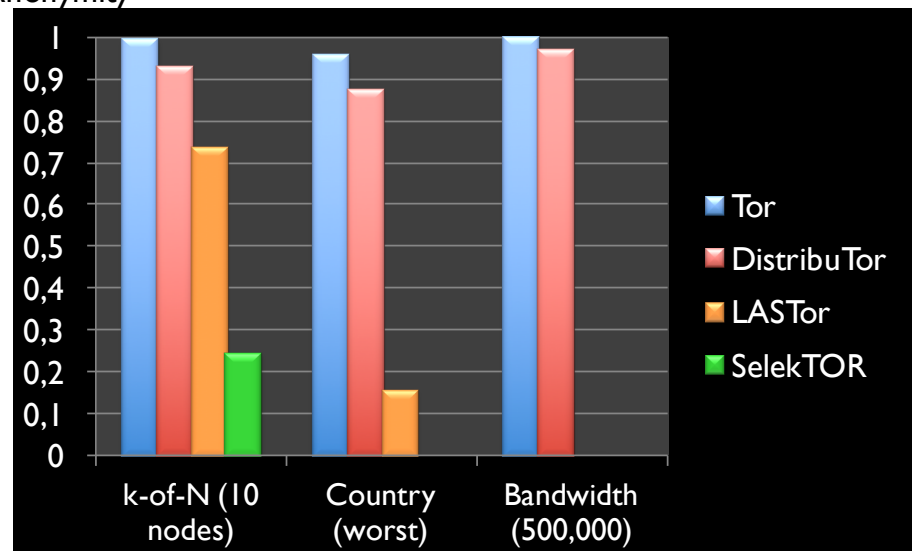
Byzantine Adversary (Relationship Anonymity)



Conversion Phase:

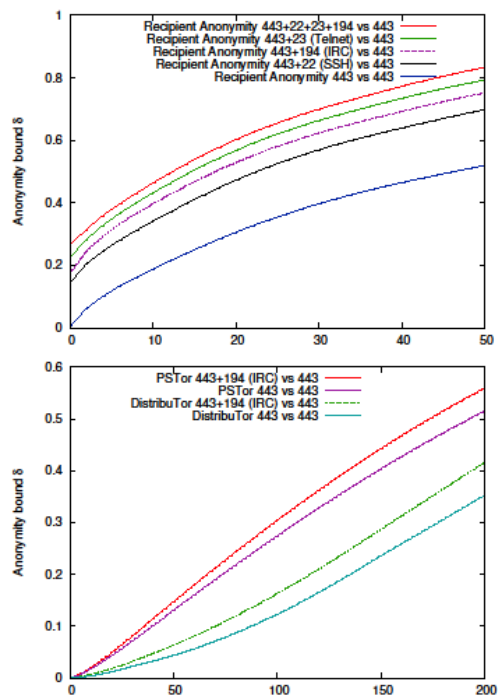
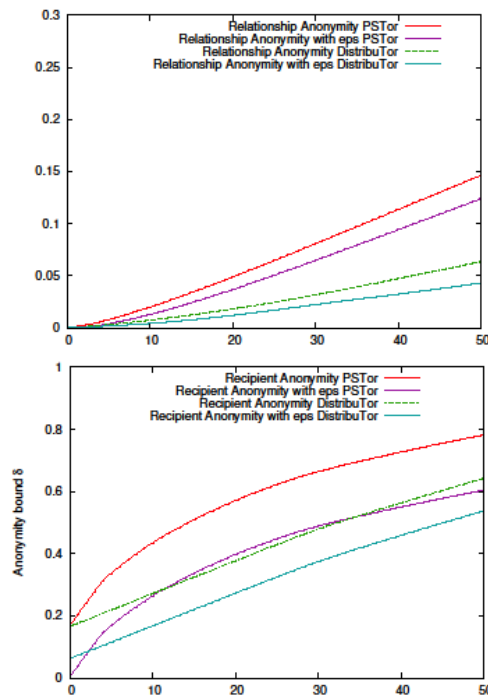
Using a Non-Standard Path Selection Algorithm

Anonymity



Scenarios:

- Tor's Path selection (**Tor**) (for comparison)
- Using DistribuTor when everyone else uses Tor's path selection (**DistribuTor**)
- Using LASTor when everyone else uses Tor's path selection (**LASTor**)
- Using SelekTOR [only US exit] when everyone else uses Tor's path selection (**SelekTOR**)



Semantic Linkability/ Content-based Anonymity

2015: User-Centric Internet

- 1) Complex trust relationships
 - Mobile devices with sensing capabilities
 - Third party software/apps & services

Vastly increased attack surface

- 2) Unprecedented dissemination of personal information
 - Advent of social networks & digital capture
 - Targeted advertisement

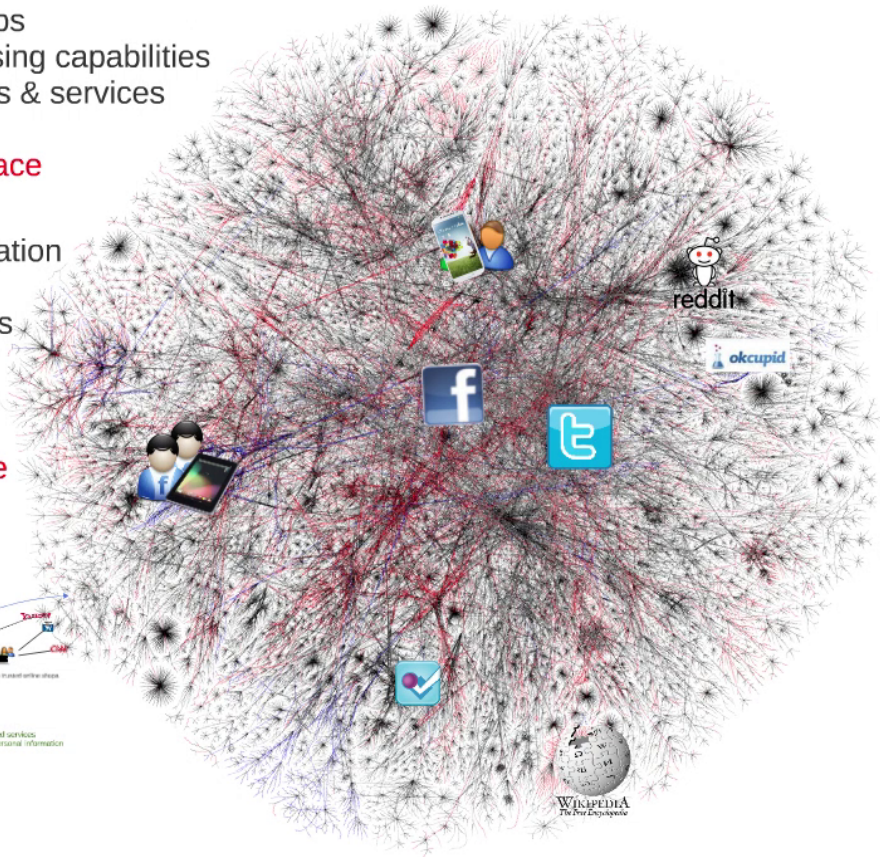
A deluge of privacy-sensitive information collected

2000: Business Internet
Fast Internet for business-to-business and business-to-consumer

Millions of online users interact with trusted online shops

Assumptions at that time:

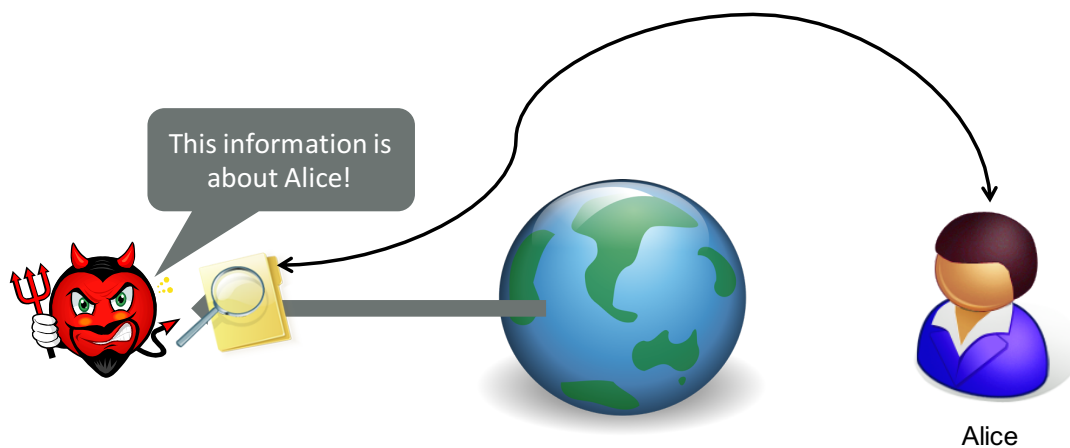
- 1) Trusted computer connects to trusted services
- 2) Limited & controlled disclosure of personal information



Definitions – Identity Disclosure

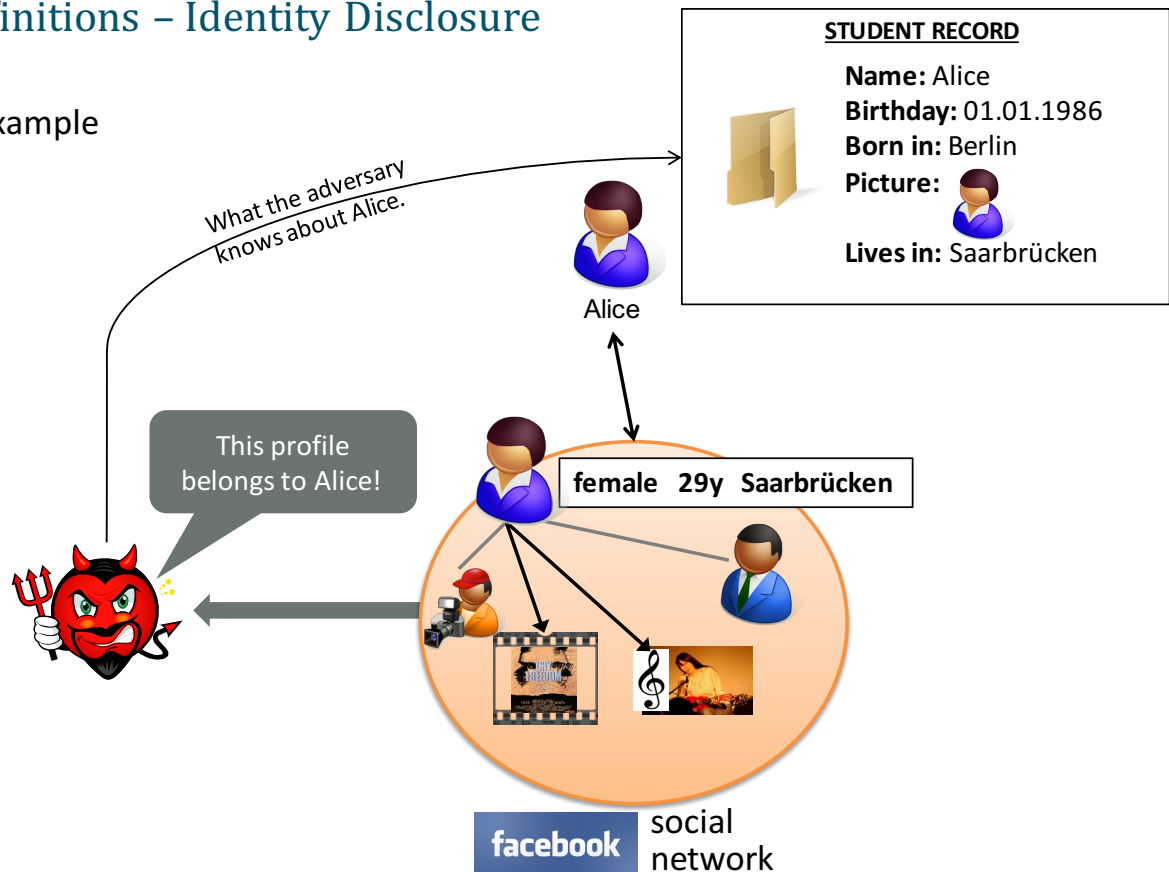
- *Identity disclosure* occurs when an adversary is able to determine the mapping from a profile v in the social network to a specific real-world entity p .

PRIVACY IN SOCIAL NETWORKS: A SURVEY (2011)



Definitions – Identity Disclosure

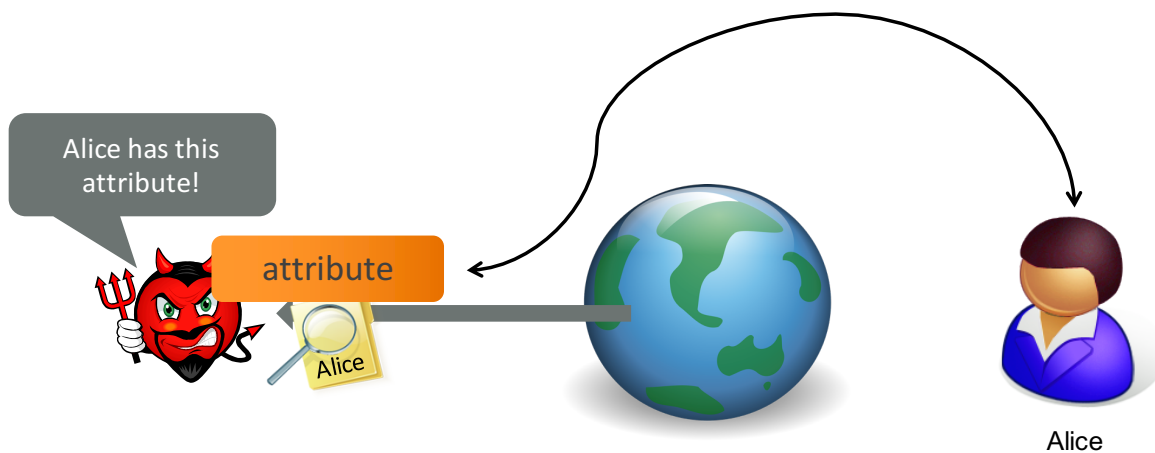
- Example



Definitions – Attribute Disclosure

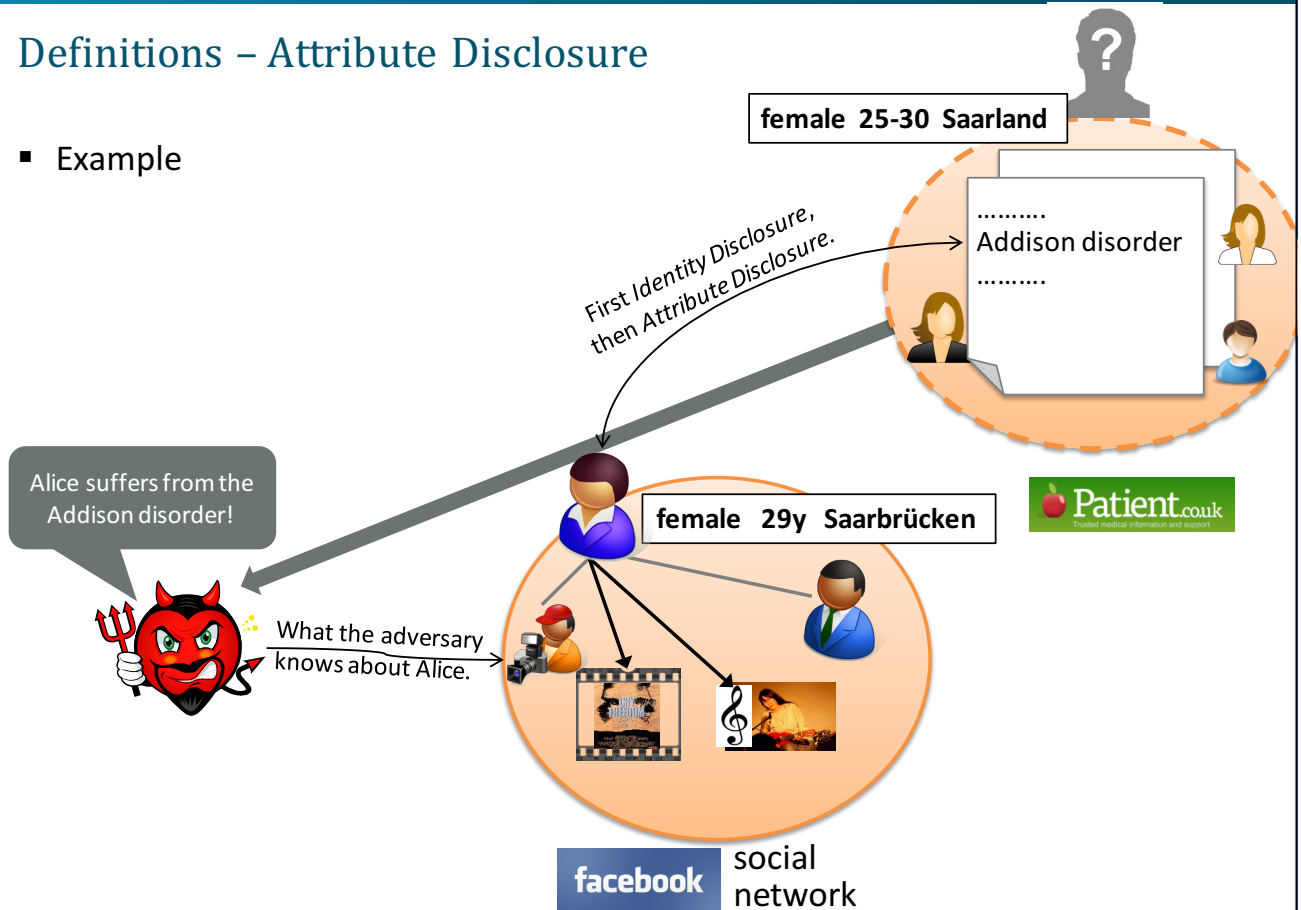
- Attribute disclosure* occurs when an adversary is able to determine the value of a sensitive user attribute, one that the user intended to stay private.

PRIVACY IN SOCIAL NETWORKS: A SURVEY (2011)



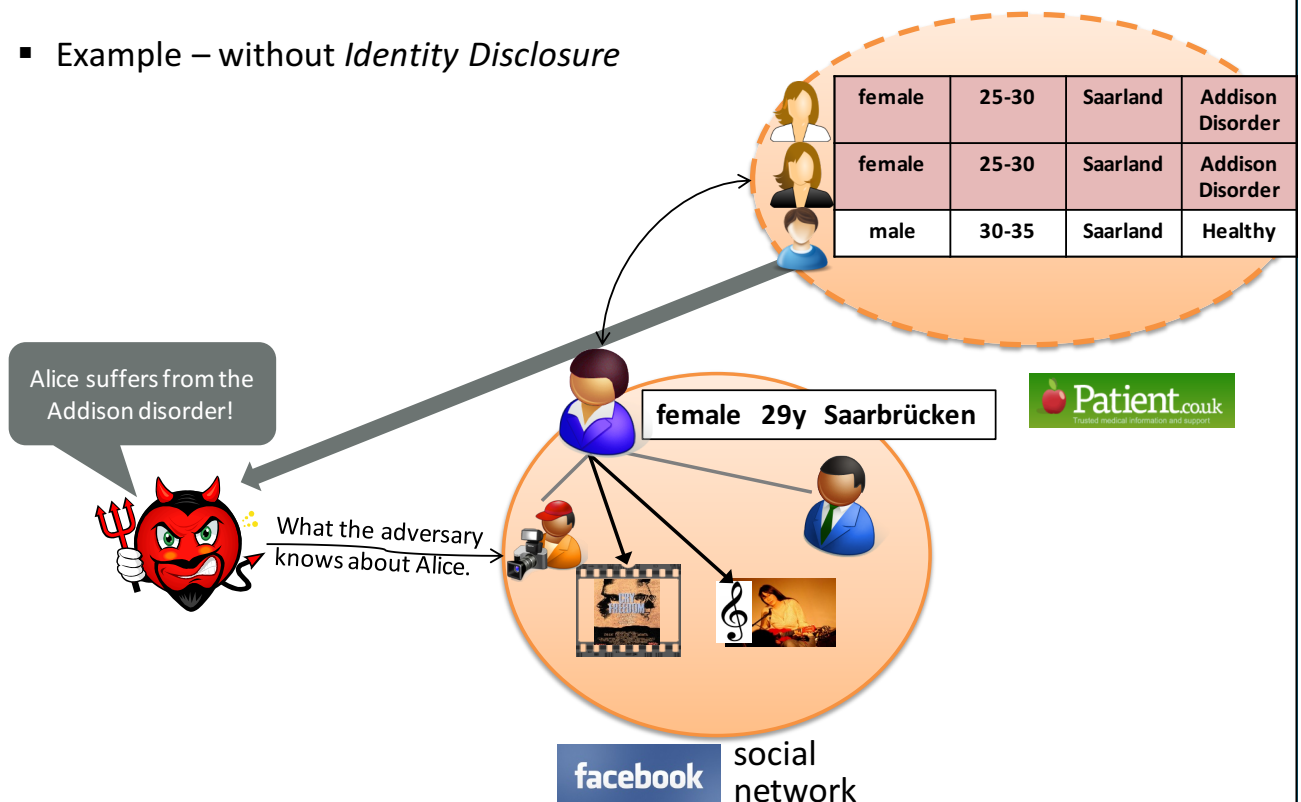
Definitions – Attribute Disclosure

- Example



Definitions – Attribute Disclosure

- Example – without *Identity Disclosure*



Database Privacy

e.g. Global Company Database
 Disproportionally Gender

ID	Zip	Age	Gender	Salary
24	61045	26	W	120000
34	67834	34	M	40000
28	12365	47	W	60000
56	24654	41	M	180000
97	98034	32	M	55000
102	12534	29	W	140000

- Structured Data
- Differentiation between Key- and Sensitive Attributes
- various privacy notions that guarantee some kind of privacy for the whole dataset
 - k-anonymity
 - l-diversity
 - t-closeness

- But which are the sensitive attributes?
- Is this the only data we can access?

-> Literature has shown that Privacy is much more diverse a problem (e.g. Netflix Challenge)

Netflix Challenge (Narayanan and Shmatikov, S&P08)

- Pre-defining a set of „sensitive“ attributes does not make sense!

Netflix Challenge:

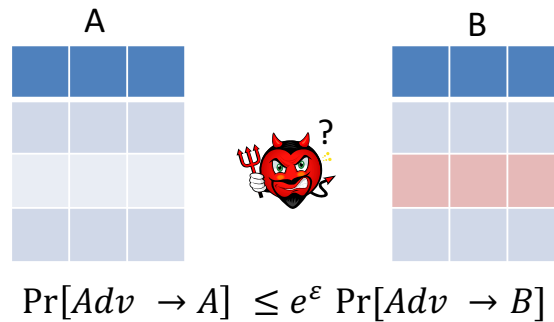
- Given: anonymized data-set of profiles with movie reviews
- Goal: identify anonymized profiles with live netflix data
- Use
 - movie reviews and
 - sparseness of data
 to identify profiles!

}

Movie reviews are sensitive data!
 Sparseness supplies auxiliary information!

Differential Privacy – for statistical databases

- Idea: Add noise to database to protect user data!



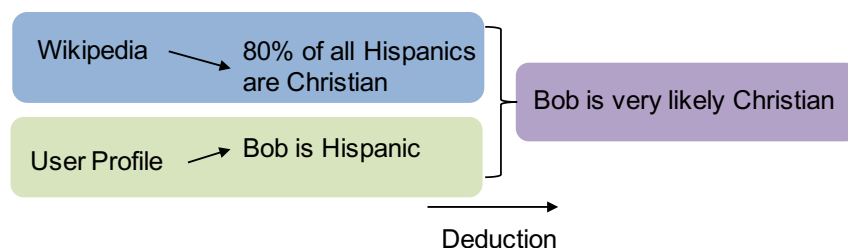
- Still restricted to structured and static data
- utility – privacy tradeoff unclear

Database Privacy vs Big-Data Privacy

Statistical Databases	Open Web/Big-Data
<ul style="list-style-type: none">▪ static and structured data▪ key- and sensitive attributes▪ mostly no adversarial background knowledge▪ privacy for the whole dataset	<ul style="list-style-type: none">▪ dynamic, heterogeneous and unstructured data▪ all information is potentially sensitive▪ ubiquitous background knowledge▪ whole dataset not known -> user centric privacy

Privacy in the open Web

- Evaluate the probability of unintended information leakage on the web
- Core Issues:
 - **Sensitivity of Information depends on context**
(e.g. discussing health issues in Facebook vs. on a health forum under an anonymous pseudonym)
 - **Unintended information disclosure through linked online profiles**
(e.g. linking anonymous profiles on various Forums to your Facebook account)
 - **Unintended information disclosure through inference**



Overview

General Linkability Model

- Linkability of Online Profiles
- d-convergence

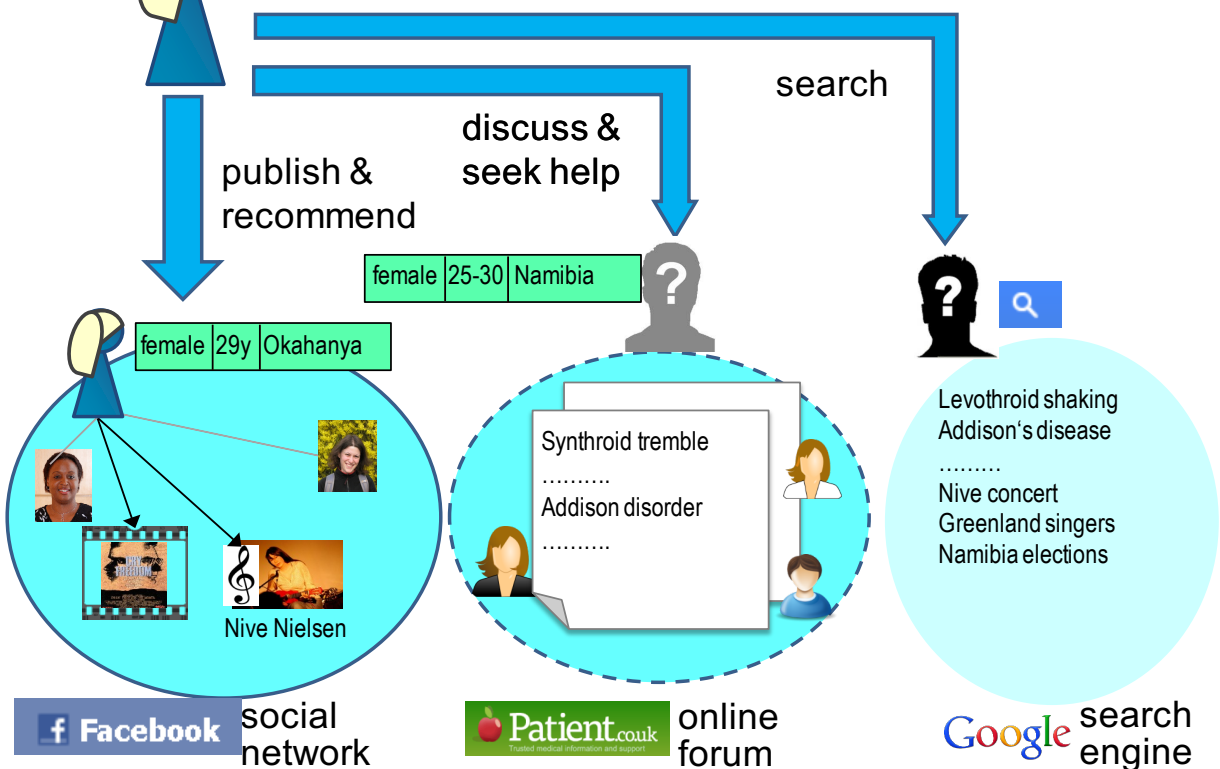
Measuring Anonymity using d-convergence

- Anonymity Estimation
- Experimental Evaluation

Authorship Attribution as a Linkability Problem

- Intro to Stylometry
- Model for Countermeasure Effectiveness
- Experimental Evaluation

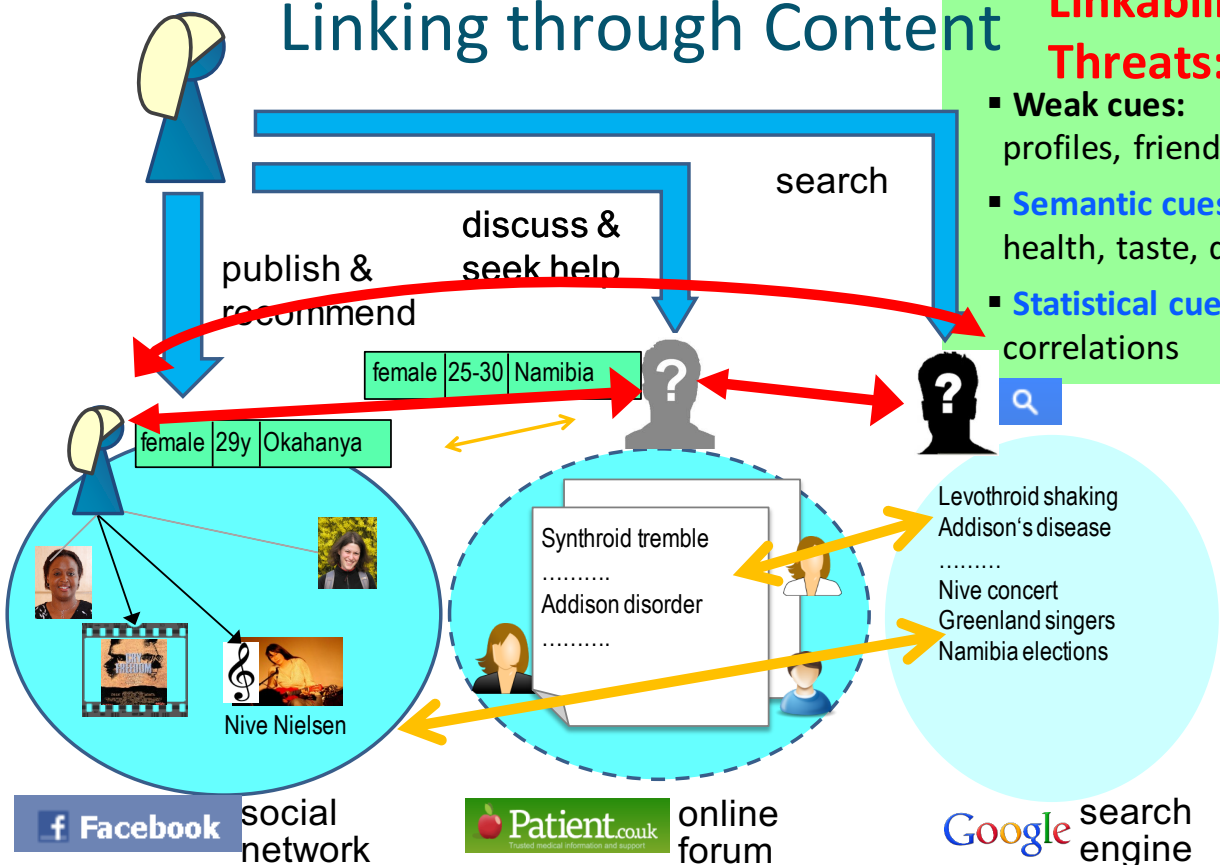
Linking through Content



Linking through Content

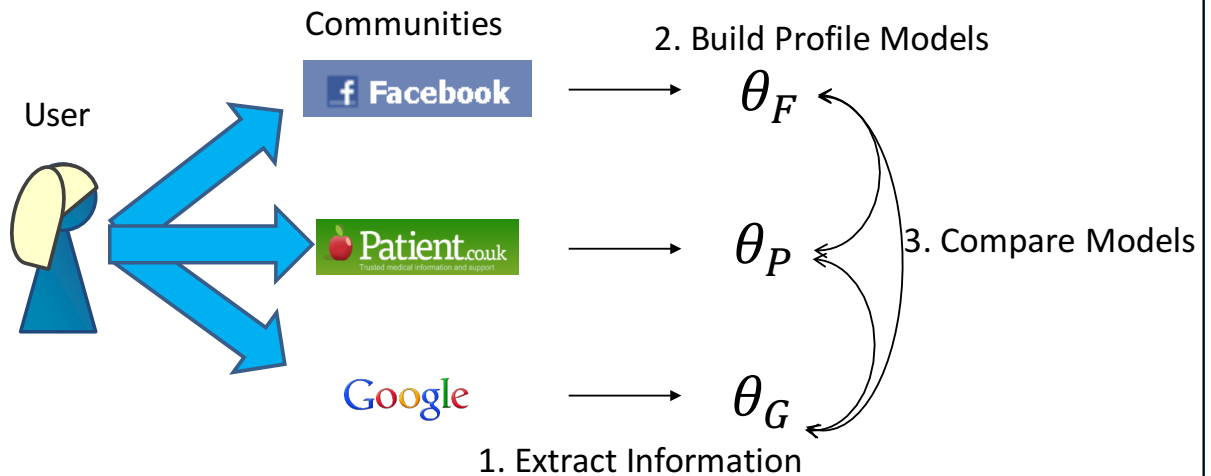
Linkability Threats:

- **Weak cues:** profiles, friends, etc.
- **Semantic cues:** health, taste, queries
- **Statistical cues:** correlations



Linking/identifying users by their content

- **Approach:** Capture Information disseminated through user content in an easily comparable manner



Statistical Model Approach to Privacy

- idea: describe entities (users, profiles, processes etc.) as statistical models
- statistical model \triangleq probability distribution over exhibited behavior

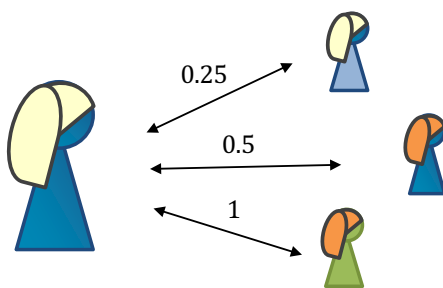
*Given a set of attributes A , the **statistical model** θ_ε of an entity ε determines the probability $\Pr[a \mid \theta_\varepsilon]$ that the entity ε exhibits attribute $a \in A$.*

Properties:

- capture arbitrary information about entities
- capture arbitrary, adversarial background knowledge by adjusting $\Pr[a \mid \theta_\varepsilon]$
- open Interpretation: $\Pr[a \mid \theta_\varepsilon]$ as probability to exhibit α as well as relevance of α to the behavior exhibited by ε

Model Similarity

- Intuition: profiles of the same user are more easily linked if their exhibited behavior is similar!
- use established distance measure $dist(\theta_1, \theta_2)$ from to define similarity



Metric:

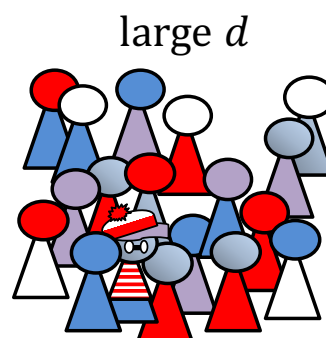
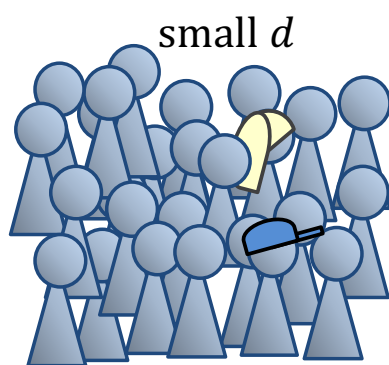
- symmetric
- triangle inequality
- $dist(\theta_1, \theta_2) \geq 0$
- If $dist(\theta_1, \theta_2) = 0$ then $\theta_1 = \theta_2$

- But: an entity is more anonymous if it behaves similarly to other entities in the same community!

d-convergence

Convergence:

- a set of entities \mathcal{E} is **d-convergent** for $\varepsilon \in \mathcal{E}$ if
$$\forall \varepsilon' \in \mathcal{E}: dist(\varepsilon, \varepsilon') \leq d$$



Interlude: Divergence Measures

- Statistical models \approx probability distributions
- Distance for probability distributions \approx **divergence**
- Popular divergence measure: **Kulback-Leibler divergence**

$$D_{KL}(\theta_1, \theta_2) = \sum_{\omega \in \Omega} \Pr[\omega \mid \theta_1] \log\left(\frac{\Pr[\omega \mid \theta_1]}{\Pr[\omega \mid \theta_2]}\right)$$

- But: **not symmetric** + some restriction on θ_1 and θ_2
- Solution: **Jensen-Shannon divergence** (symmetric variant of D_{KL})

$$D_{JS}(\theta_1, \theta_2) = \frac{1}{2} D_{KL}(\theta_1, M) + \frac{1}{2} D_{KL}(\theta_2, M)$$
$$M = \frac{1}{2} \theta_1 + \frac{1}{2} \theta_2$$

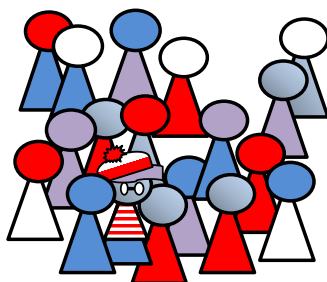
- Properties: symmetric, bounded by 1, and less restrictions

(k,d)-anonymity

- in practice, we do not know the whole set of entities \mathcal{E}
- we need to give local guarantees

*An entity ε is (k, d) – anonymous in \mathcal{E} if there is an **anonymous subet** $A_\varepsilon \subseteq \mathcal{E}$ such that*

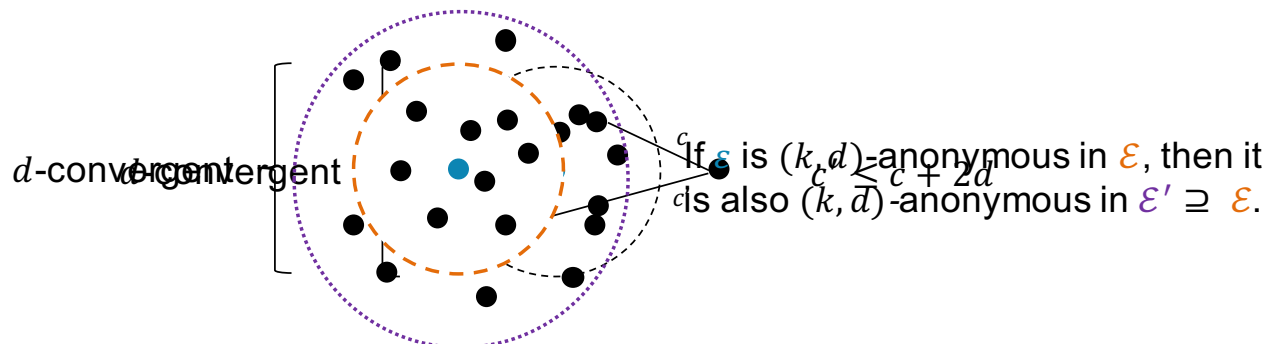
- $\varepsilon \in A_\varepsilon$
- $|A_\varepsilon| \geq k$
- A_ε is d – convergent



→ anonymous subset for $k = 6$

Properties of d-convergence

- several intuitive properties hold for convergent sets:



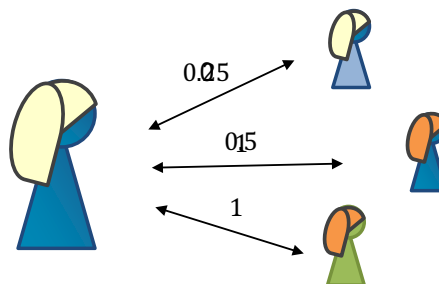
- If we restrict ourselves to the database setting, we get t-closeness:

d-convergence \Rightarrow t-closeness:

If \mathcal{E} is a d-convergent collection of entities, then $D_{\mathcal{E}}$ is d-close.

Caveat: Features and Feature Selection

- Attribute set $A \approx$ Features extracted from user's actions and content
- However:** which features does the adversary consider?



- Guarantees derived from d-convergence are only as good as the considered feature set
- Sound guarantees require best possible feature sets
 - > still open problem in Machine Learning

Overview

General Linkability Model

- Linkability of Online Profiles
- d-convergence

Measuring Anonymity using d-convergence

- Anonymity Estimation
- Experimental Evaluation

Next!

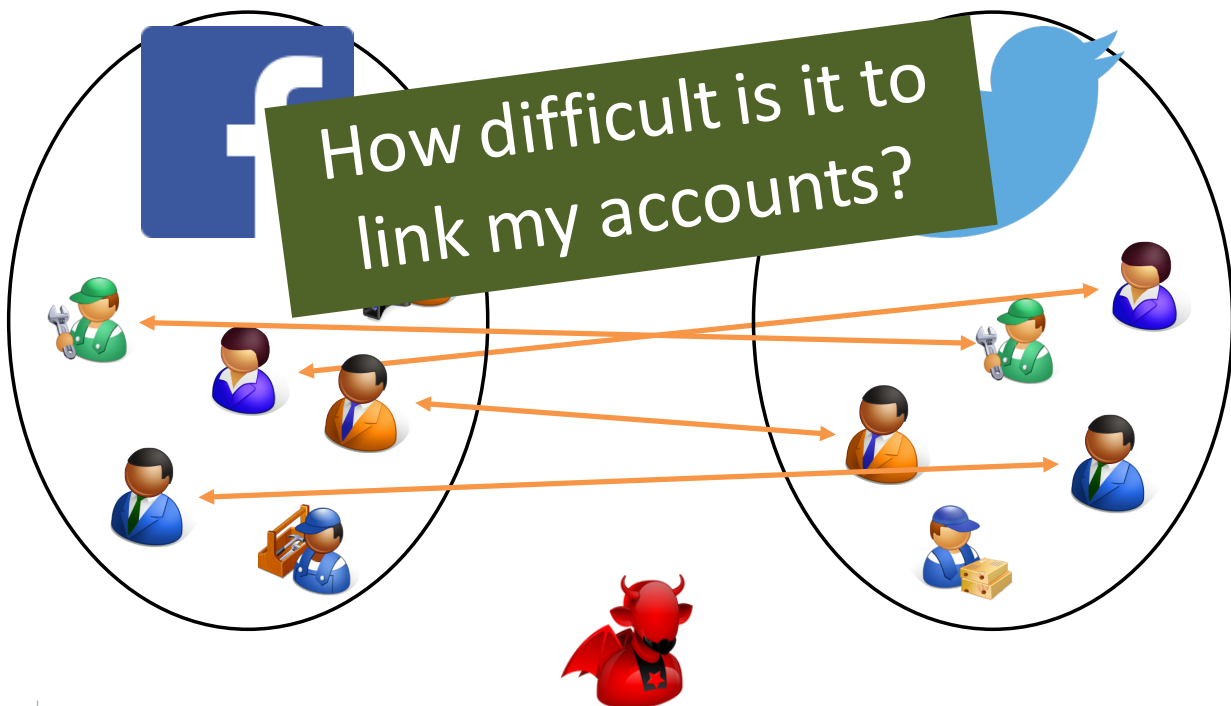
Authorship Attribution as a Linkability Problem

- Intro to Stylometry
- Model for Countermeasure Effectiveness
- Experimental Evaluation

Measuring Anonymity using d-convergence

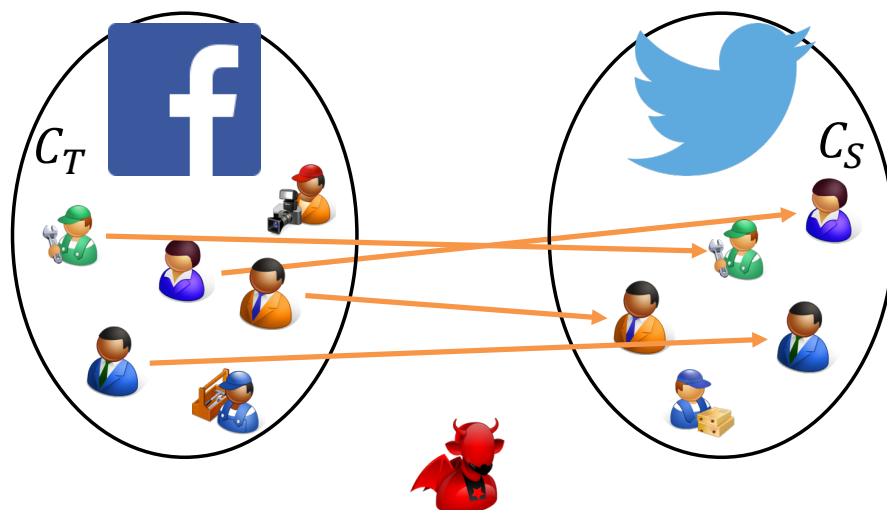
What kind of Anonymity?

- Adversary tries to **find as many matching identities as possible** across different platforms.



Adversarial Strategy – More Precisely

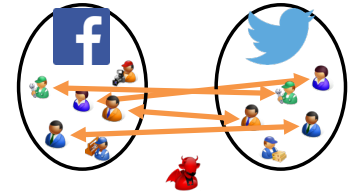
- Given a **source identity** I_S in a **source community** C_S , the adversary tries to identify the matching **target identity** I_T within a **target community** C_T that belongs to the same user.
- The adversary wants to match as many identities as possible between the two communities C_S and C_T .







Adversarial Strategy

- Adversary is *rational*:

- It matches identities based on **how similar** they are.



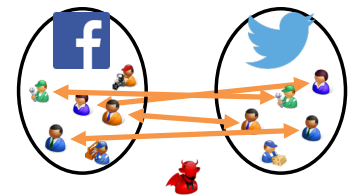
				...
	0.2	0.9	0.3	...
	0.1	0.4	0.8	...
	0.1	0.3	0.2	...
...









1. Compute pairwise **similarities** between identities in communities.
2. Compute **likelihood** that two identities belong to the same user. (based on 1.)
3. **Rank** pairs based on 2.
4. Choose a **threshold** th on the likelihood and link all identities above the threshold.

Adversarial Strategy

- Adversary is *rational*:

- It matches identities based on **how similar** they are.

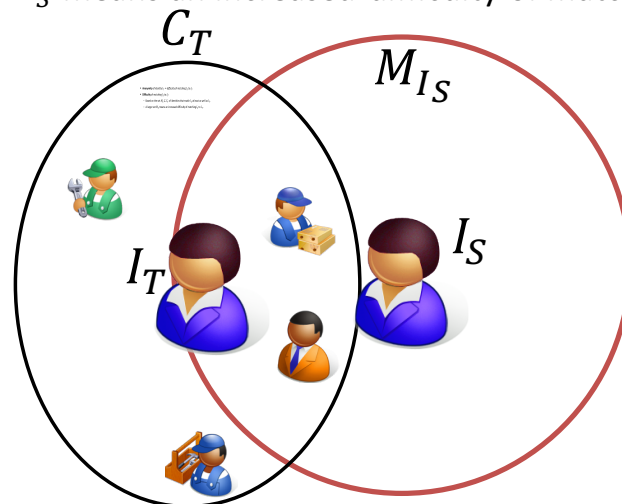


 	Likeli-hood	Rank
 	0.9	1
 	0.8	2
...
 	0.1	30

1. Compute pairwise **similarities** between identities in communities.
2. Compute **likelihood** that two identities belong to the same user. (based on 1.)
3. **Rank** pairs based on 2.
4. Choose a **threshold** th on the likelihood and link all identities above the threshold.

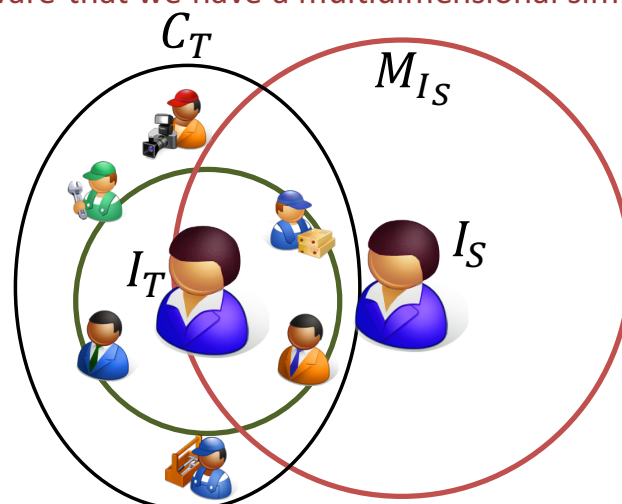
Matching Set

- **Anonymity** of identity $I_T \approx \text{difficulty}$ of matching I_S to I_T .
- **Difficulty** of matching I_S to I_T :
 - Based on the **matching set** M_{I_S} of identities that match I_S at least as well as I_T .
 - A larger set M_S means an increased difficulty of matching I_S to I_T .



Anonymous Subsets – Intuitive Hypothesis

- The *more identities are in a short distance* of I_T , the larger M_{I_S} should be.
 - This is captured by the previous definition of (k, d) -anonymous subsets.
- However, in special cases, this intuition might be **wrong**!
 - E.g., all identities in the anonymous subset are located in one part of the subset (*beware that we have a multidimensional similarity/distance!*).



Case Study on Anonymity in Reddit

- Reddit dataset consisting of
 - 15 million comments
 - 58.000 identities, 37.000 users
 - 1.930 subreddits
- Automatically provides ground truth between subreddits by user pseudonyms
- Graphs generated for $C_S = \text{news}$ and $C_T = \text{worldnews}$
- Distance based on *word unigrams*
- 3 specific identities are highlighted for further inspection



Relative vs. Absolute Anonymity Measures

- How well can we estimate the *difficulty* (size of M_{I_S}) by anonymous subsets?
 - It depends on the *information available*
 - and thus on the party *who* is estimating it.

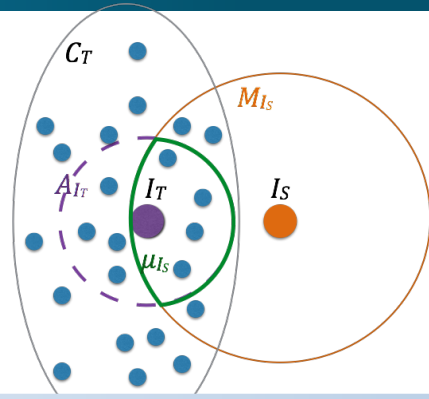
The User Herself	The Service Provider (e.g., Twitter)
knows her own matching accounts	only knows identities in his system
limited computational power	much larger computational power

Can estimate her
absolute anonymity.

Can estimate only the **relative** anonymity,
e.g., which identity is more at risk.

Absolute Anonymity Measure

- Given the matching identities I_S and I_T ,
- how can we provide a **lower bound** on the matching set M_{I_S} ?



Let $d = \text{dist}(\theta_{I_S}, \theta_{I_T})$. Then the **local matching set** μ_{I_S} of the source identity I_S matching against a target identity I_T is defined by

$$\mu_{I_S} = M_{I_S} \cap A_{I_T}(k, d).$$

- Clearly, $\mu_{I_S} \subseteq M_{I_S}$ provides a lower bound on M_{I_S} .
- Local matching set $\mu_{I_S} = M_{I_S} \cap A_{I_T}(k, d)$ can be computed easily **without** considering all identities in C_T .

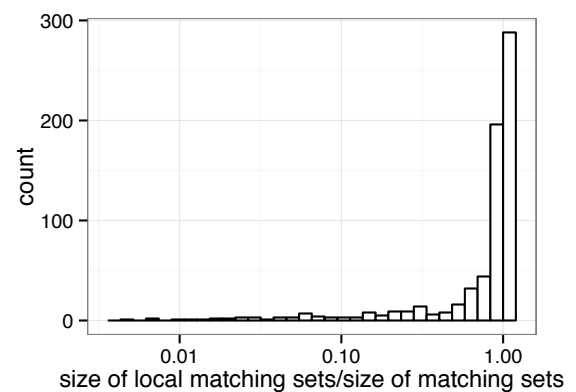
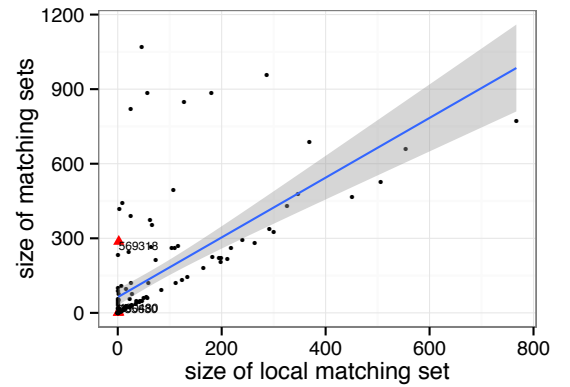
Local Matching Sets

Algorithm:

1. Compute A_{I_T} for $d = \text{dist}(\theta_{I_S}, \theta_{I_T})$ (we do not need to compute the largest anonymous subset for that d).
2. Compute the distance $\text{dist}(\theta_{I_S}, \theta_{I'})$ between I_S and every $I' \in A_{I_T}$.
3. If $\text{dist}(\theta_{I_S}, \theta_{I'}) \leq \text{dist}(\theta_{I_S}, \theta_{I_T})$, then $I' \in \mu_{I_S}$.
4. Incrementally increase $|A_{I_T}|$ by repeating the process until we reach an anonymity threshold with which we are comfortable.

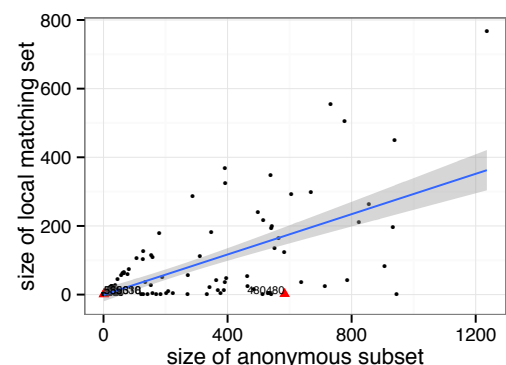
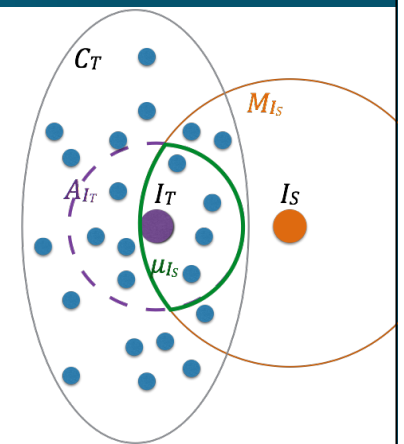
Local Matching Sets

- Indeed, the size of the local matching set $|\mu_{I_S}|$ is a good approximation for the matching set M_{I_S} .
- Analyzing the ratio $|\mu_{I_S}|/|M_{I_S}|$ over all subreddits shows that:
 - In $\geq 74\%$ of the cases, $|\mu_{I_S}| \geq 0.8 \cdot |M_{I_S}|$



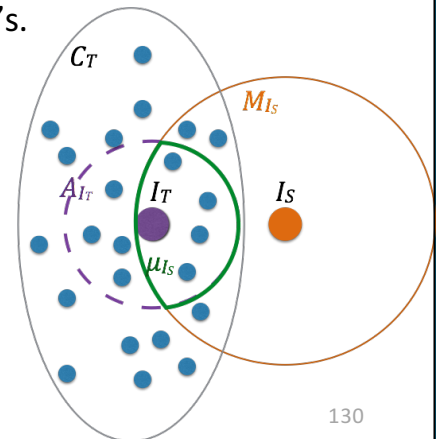
Approximating Local Matching Sets

- Local matching set $\mu_{I_S} = M_{I_S} \cap A_{I_T}(k, d)$ can be computed easily **without** considering all identities in C_T .
- However, computing $\text{dist}(\theta_{I_S}, \theta_{I'})$ between I_S and every $I' \in A_{I_T}$ may be **too expensive!**
- Solution: If A_{I_T} is *evenly distributed*, it can serve as an approximation of μ_{I_S} .
 - In our case study, the size of A_{I_T} correlates with the size of μ_{I_S} .
 - By definition $|\mu_{I_S}| \leq |A_{I_T}|$.
 - A **few** outliers that underapproximate the risk, because A_{I_T} is not evenly distributed (see later).



Relative Anonymity Measure

- Given only the identities in C_T ,
- how can we provide a **relative ranking** on the matching sets $M_{I'}$ for $I' \in C_T$?
- Idea: Use anonymous subsets $A_{I'}$ for approximation.
- Question: What convergence d should be used to compute $A_{I'}$?
 - If the specific matching strategy implies a certain value for d , we are fine.
 - Otherwise: compute rank across many different d 's.

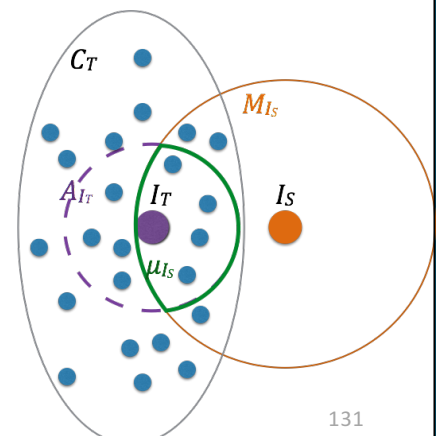


Relative Anonymity Measure – Ranking for given d

- Assumption: Given a fixed value for d .

Algorithm:

1. Compute the maximum $(k_{I'}, d)$ -anonymous subsets $A_{I'}$.
2. **Rank** all identities in C_T by their **anonymous subset size** $k_{I'} = |A_{I'}|$.






Relative Anonymity Measure – Ranking

- Assumption: No specific value for d is given.

Algorithm:

- Choose different values for d , and **rank** all identities in C_T according to each d as if d was given.
 - Resolve ties by assigning identities all possible ranks they could occupy.

Identity	$ A_{I'} $	Rank
 I_1	2	1
 I_2	5	2, 3
 I_3	5	2, 3

- Compute a **global ranking** from rankings for different values of d .

Relative Anonymity Measure – Ranking




- Assumption: No specific value for d is given.

Algorithm:

- Choose different values for d , and **rank** all identities in C_T according to each d as if d was given.
- Compute a **global ranking** from rankings for different values of d .
 - Construct bipartite graph between identities and their ranks for different d 's.
 - The weight of an edge (I', i) corresponds to the number of times I' was ranked at position i .
 - The global ranking is then given by the **maximum weight matching**.

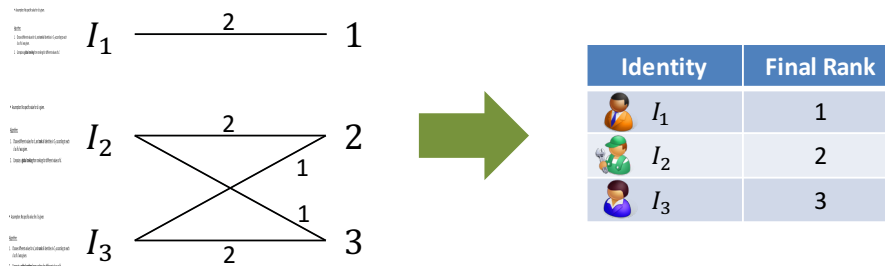
Relative Anonymity Measure – Ranking

- Assumption: No specific value for d is given.

Identity	Rank d_1	Rank d_2
 I_1	1	1
 I_2	2, 3	2
 I_3	2, 3	3

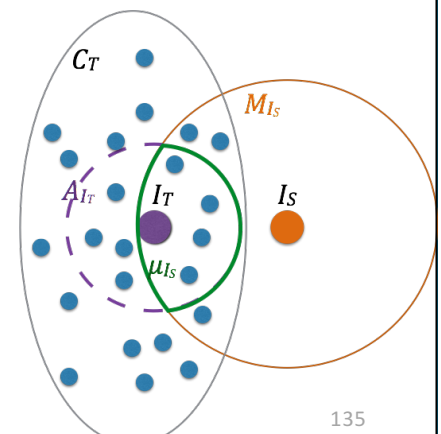
Algorithm:

- Choose different values for d , and **rank** all identities in C_T according to each d as if d was given.
- Compute a **global ranking** from rankings for different values of d .



Relative Anonymity Measure

- Given only the identities in C_T ,
- how can we provide a **relative ranking** on the matching sets $M_{I'}$ for $I' \in C_T$?
- Ranking by anonymous subsets sizes $|A_{I'}|$.
- Indicates which identities are more at risk:
 - Larger numeric rank** means larger anonymous subset
 - and most likely a larger matching set (we will show that empirically).
 - It is more difficult to correctly link those identities compared to identities with a smaller rank.

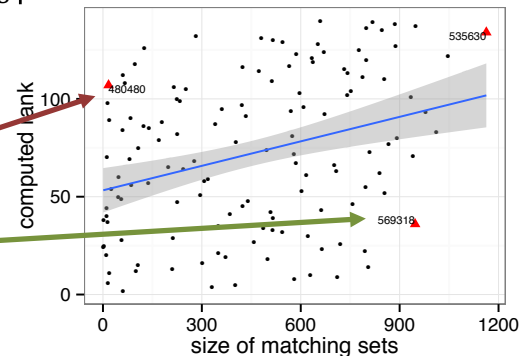


Relative Anonymity Measure

- Expectation: Correlation between rank and $|M_{IS}|$.

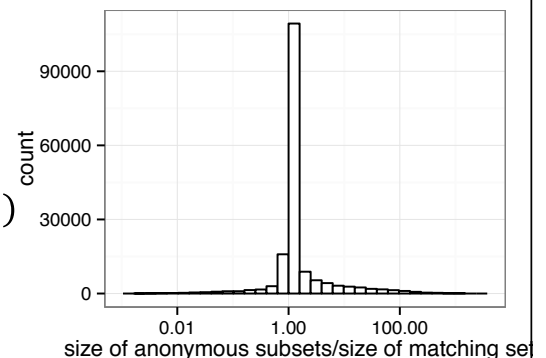
- Correlation exists, but there are outliers:

- that underapproximate the risk,
- that overapproximate the risk.



- Outliers are minority as $|A_{IT}|/|M_{IS}|$ suggest.

- overapproximation results in ratio < 1
- underapproximation results in ratio > 1
- In $\geq 71\%$ of all cases, the ratio is $\in [0.8, 1.2]$



Ranking – An Adversary's Tool

- Given the **ranking** of identities by their anonymous subset sizes,
- an adversary can use it to improve its *precision* by **only matching** those identities with small anonymous subsets.
- Intuition: Identities with larger anonymous subsets are more difficult to link.
- **If the adversary does not even try linking these, the process yields less false positive matchings.**

Overview

General Linkability Model

- Linkability of Online Profiles
- d-convergence

Measuring Anonymity using d-convergence

- Anonymity Estimation
- Experimental Evaluation

Authorship Attribution as a Linkability Problem

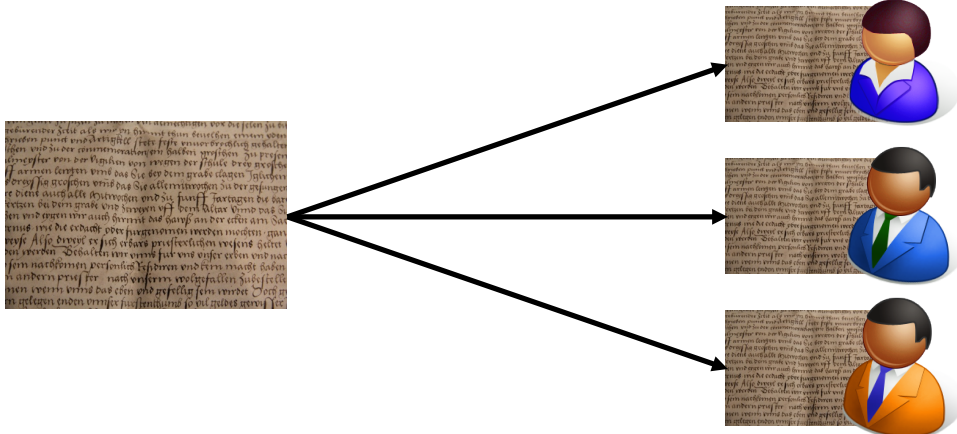
- Intro to Stylometry
- Model for Countermeasure Effectiveness
- Experimental Evaluation

Next!

Authorship Attribution as a Linkability Problem

What is Authorship Attribution?

- Given
 1. a text written by an *anonymous/unknown author*
 2. and texts of *known authors*.
- **Who** was the author of the 1. text?
- Analyses the 'writing style' of the authors.



Authorship Attribution as a Linkability Problem

Stylometry

- Also called *Writing Style* or *Writeprint*.
- Analyses the linguistic style, usually of written language.
- Can isolate **rare elements of a text** or even **identifying patterns in common parts of speech**.
- Authorship Attribution uses the **authors' stylometry** to distinguish between the different authors.

Stylometric Features

- Stylometry is determined by analysing an author's texts with respect to **stylometric features**.
- A stylometric feature captures one **linguistic characteristic** of the text.
- Examples for stylometric features are:
 - *Number of words*, e.g., 10
 - Frequencies of different *letters*, e.g., 3 times a
 - *Word unigrams/bigrams/trigrams*, e.g., 5 times house
 - Frequency of *Part of Speech tags*, e.g., 4 times NP (nominal phrase)
 - Frequency of *Misspelled Words*, e.g., 8 times muose

Stylometric Features - Writeprints

- Famous set of stylometric features called **Writeprints** [Abbasi and Chen 2008]
- has been used for authorship recognition of known authors
- and similarity detection of unknown authored documents.

Group	Category	Quantity		Description
		Baseline (BF)	Extended (EF)	
Lexical	Word-Level	5	5	total words, % char. per word
	Character-Level	5	5	total char., % char. per message
	Letters	26	26	count of letters (e.g., a, b, c)
	Character Bigrams	—	<676	letter bigrams (e.g., aa, ab, ac)
	Character Trigrams	—	<17,576	letter trigrams (e.g., aaa, aab, aac)
	Digits	—	10	digits (e.g., 1, 2, 3)
	Digit Bigrams	—	<100	2 digit number frequencies (e.g., 10, 11)
	Digit Trigrams	—	<1,000	frequency of 3 digit numbers (e.g., 100)
	Word Length Dist.	20	20	frequency of 1–20 letter words
	Vocab. Richness	8	8	richness (e.g., hapax legomena, Yule's K)
	Special Characters	21	21	occurrence of special char. (e.g., @#%`)
Syntactic	Function Words	150	300	frequency of function words (e.g., of, for)
	Punctuation	8	8	occurrence of punctuation (e.g., !,;,?)
	POS Tags	—	<2,300	frequency of POS tags (e.g., NP, JJ)
	POS Tag Bigrams	—	varies	POS tag bigrams (e.g., NP VB)
	POS Tag Trigrams	—	varies	POS tag trigrams (e.g., VB JJ)
	Message-Level	6	6	e.g., has greeting, has url, quoted content
	Paragraph-Level	8	8	e.g., no. of paragraphs, paragraph lengths
	Technical Structure	50	50	e.g., file extensions, fonts, use of images
	Words	20	varies	bag-of-words (e.g., "senior", "editor")
	Word Bigrams	—	varies	word bigrams (e.g., "senior editor")
Content	Word Trigrams	—	varies	word trigrams (e.g., "editor in chief")
	Misspelled Words	—	<5,513	misspellings (e.g., "beleive", "though")
Idiosyncratic		—	<5,513	misspellings (e.g., "beleive", "though")

Authorship Attribution

- Most often:
 - Train classifier (e.g., SVM) on corpus of texts of known authors (using a set of stylometric features)
 - Use classifier to find author of unknown authored texts
- Applications of Authorship Attribution:
 - Literature (e.g., classify epochs, identify author of documents)
 - Forensics
- Performance of current methods:
 - Writeprints [Abbasi and Chen 2008]: 100 authors, 94% accuracy
 - [Narayanan et al. 2012]: 100.000 authors, 20% accuracy

Authorship Attribution – Quotes

- Former Wikileaks spokesman: “If someone had run WikiLeaks documents through such a program, he would have discovered that the same two people were behind all the various press releases, document summaries, and correspondence issued by the project. The official number of volunteers we had was also, to put it mildly, grotesquely exaggerated.” [Domscheit-Berg et al. 2011]
- FBI report: “As non-handwritten communications become more prevalent, such as blogging, text messaging and emails, there is a growing need to identify writers not by their written script, but by analysis of the typed content. Currently, there are some studies in the area of writer’s colloquial analysis that may lead to the emerging technology of writer identification in the blogosphere.” [Colosimo et al. 2009]

Adversarial Stylometry

- Privacy perspective on authorship attribution:
 - Is it possible to **circumvent authorship attribution**?
 - Authorship attribution, in principle, can deanonymize activists, journalists, bloggers,...
- [Brennan et al. 2012] tested the impact of several methods to circumvent authorship attribution:
 - Manual obfuscation **is possible**
 - Imitation **is possible**
 - Machine Translation (e.g., English → German → English) **does not help**

Adversarial Stylometry

- Manual obfuscation **is possible** but **hard**
- [McDonald et al. 2012] presented a semi-automated framework called **Anonymouth**:
 - Analyses stylometric features and gives **hints** what to change
 - E.g., *use fewer instances of the letter “i”*
 - Clearly, such hints can be hard to realize
- Open questions:
 - Need for **automated** techniques!
 - How to **formally evaluate** the effect of adversarial stylometry in detail?

Applying d -convergence

- We already have a notion of distance between statistical models for, e.g., *word unigrams*.
- The **similarity** of two texts (or collections of texts) with respect to a class of stylometric features τ (e.g., word unigrams) is given by
$$\text{sim}(\theta_{T_1}^\tau, \theta_{T_2}^\tau) = 1 - \text{dist}(\theta_{T_1}^\tau, \theta_{T_2}^\tau).$$
- The **statistical model** of an identity consists of the statistical models of all types of stylometric features $\theta_I = (\theta_{\tau_1}, \dots, \theta_{\tau_n})$.

The **similarity** of two identities is then defined as a linear combination:

$$\text{sim}(\theta_{I_1}, \theta_{I_2}) = \lambda_1 \cdot \text{sim}(\theta_{I_1}^{\tau_1}, \theta_{I_2}^{\tau_1}) + \dots + \lambda_n \cdot \text{sim}(\theta_{I_1}^{\tau_n}, \theta_{I_2}^{\tau_n}) + \rho$$

- Unknown values $\lambda = (\lambda_1, \dots, \lambda_n)$ and ρ will be determined later.

Applying d -convergence

- Unknown values $\lambda = (\lambda_1, \dots, \lambda_n)$ and ρ are obtained by **training a classifier**.
- $A = \begin{pmatrix} \text{sim}(\theta_{I_1}^{\tau_1}, \theta_{I_2}^{\tau_1}) & \dots & \text{sim}(\theta_{I_1}^{\tau_n}, \theta_{I_2}^{\tau_n}) \\ \vdots & \ddots & \vdots \\ \text{sim}(\theta_{I_j}^{\tau_1}, \theta_{I_k}^{\tau_1}) & \dots & \text{sim}(\theta_{I_j}^{\tau_n}, \theta_{I_k}^{\tau_n}) \end{pmatrix}$, \mathbf{b} with $b_i = \begin{cases} 1 & \text{if same author} \\ 0 & \text{otherwise} \end{cases}$
- Examples:
 - Least Squares
Minimizes $\|A\lambda - \mathbf{b}\|_2^2$. To get ρ , a special 1 column has to be added to A .
 - Linear SVM with class labels \mathbf{b}
Returns decision function $D(\mathbf{x}) = \lambda \cdot \mathbf{x} + \rho$
(in this case: $\text{sim}(\theta_{I_1}, \theta_{I_2}) = D\left((\text{sim}(\theta_{I_1}^{\tau_1}, \theta_{I_2}^{\tau_1}) \dots \text{sim}(\theta_{I_1}^{\tau_n}, \theta_{I_2}^{\tau_n}))^T\right)$).
- Similarity can then be used to attribute authorship.

Importance of a Feature

- How to **formally evaluate** the effect of adversarial stylometry in detail?
 - Ideally, we want to reduce the **importance** of identifying features.
- Information Gain
 - $IG(F_i) = H(I) - H(I | F_i)$, with Shannon-Entropy H , feature F_i , identity I
 - Intuitively, $IG(F_i)$ is higher if F_i is *more discriminating*
 - Classifier-independent, but
 - does not take the actual matching into account
(a feature may be discriminating within one community, but not both)
- Precision/Recall/Accuracy of classifier
 - Gives a global assessment of the effect, but fails in providing importance of particular features

Importance of a Feature

- How to **formally evaluate** the effect of adversarial stylometry in detail?
 - Ideally, we want to reduce the **importance** of identifying features.
- Weights of $\text{sim}(\theta_{I_1}, \theta_{I_2}) = \lambda_1 \cdot \text{sim}(\theta_{I_1}^{\tau_1}, \theta_{I_2}^{\tau_1}) + \dots + \lambda_n \cdot \text{sim}(\theta_{I_1}^{\tau_n}, \theta_{I_2}^{\tau_n}) + \rho$
 - Gives a feature-class-level assessment for the optimal matching
 - Has been successfully applied in other areas (Gene selection [Guyon et al. 2002])

The **importance** of the feature class τ_i in terms of the optimal matching is defined as:

$$\text{imp}(\tau_i) = (\lambda_i)^2$$

Gain – The Effect of Adversarial Stylometry

- On a high level, **gain** is the *difference* between some importance or performance assessment *before and after an obfuscation*.
- In terms of the **importance** of a feature-class, we have λ_i before and λ'_i after the obfuscation:

The **gain** of an obfuscation w.r.t. the feature class τ_i is given by:

$$\text{gain}(\tau_i) = (\lambda_i)^2 - (\lambda'_i)^2$$

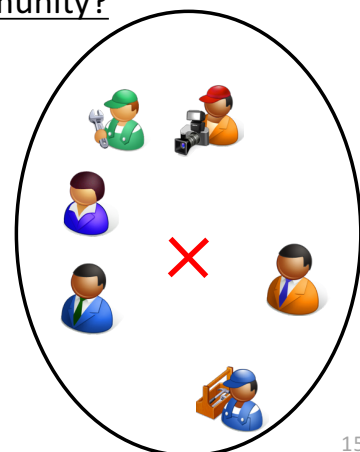
- Implicitly, we assume that a feature's importance of 0 is the best value an obfuscation can achieve (the feature is not important at all anymore).
- The **overall gain** is then given by $\text{gain} = \sum_i \text{gain}(\tau_i)$.

Fully Automated Obfuscations – A Case Study





- More generally: Algorithms that automatically modify text.
 - Case study on:
 - Spellchecker
 - Also corrects grammar
 - Synonym replacement tool
 - E.g., highly → extremely
 - Tool for **adding and removing** misspellings
 - E.g., irrelevant → irllevant
 - Tool for replacing special characters by their meaning and vice versa
 - E.g., © → copyright
- * If there are multiple alternatives, which should be chosen?

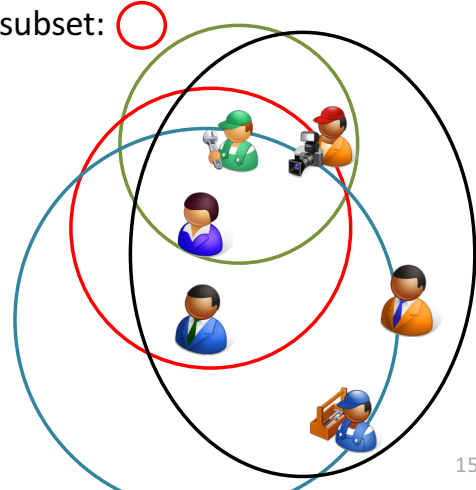
Multiple Alternatives

- If an automated text modification has to choose between multiple actions, how to decide?
- **Goal:** Make text *more similar* to those of *other authors*. Then, identification should be intuitively harder.
- Try to get close to the “average identity” in the community?
- Might result in being more identifying!
 - See right, average identity **✗** is far away from actual identities.



Multiple Alternatives – Use Anonymous Subsets

- If an automated text modification has to choose between multiple actions, how to decide?
 - **Goal:** Make text *more similar* to those of *other authors*. Then, identification should be intuitively harder.
1. For all identities I within the k -anonymous subset: 
 - Determine the convergence of I 's k -anonymous subset.  
 2. Choose action that makes text closer to I with minimal convergence. 
- **Intuition:** Find next identity with many other identities around.



Methodology for Evaluating Effectiveness

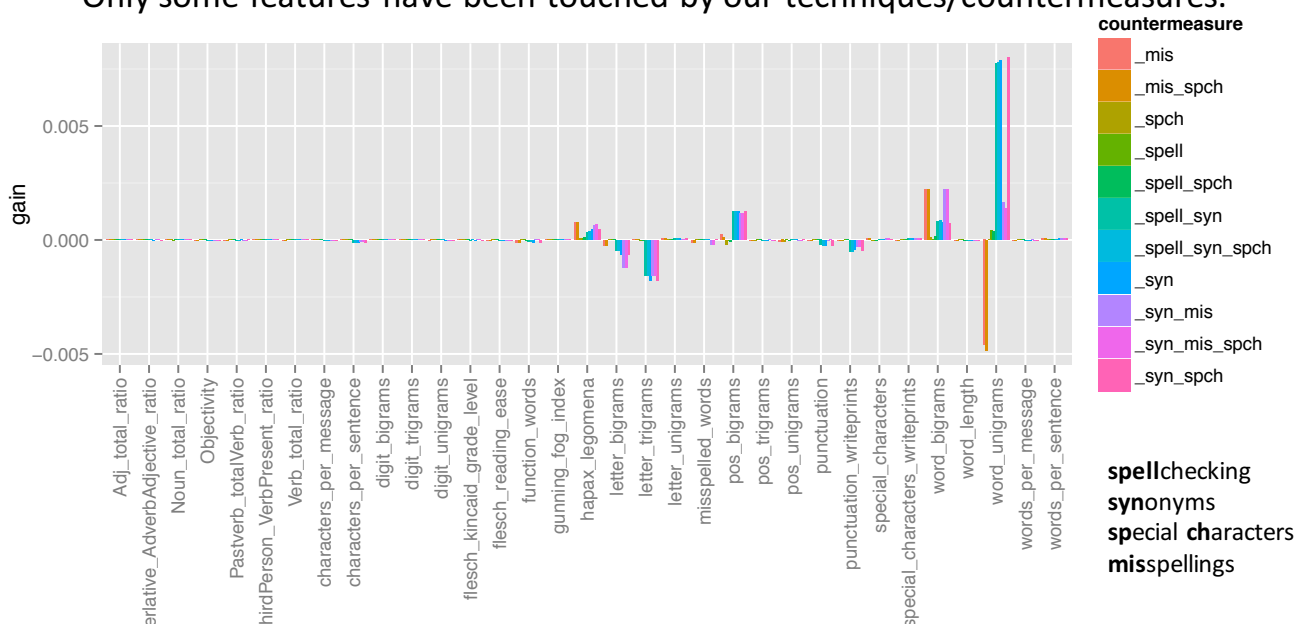
1. Compute features and similarities on texts.
2. Train classifier to obtain λ and ρ .
3. Apply obfuscation techniques on a set of **test users**, each applying the obfuscation individually (do not model interdependencies).
 - Also apply combinations thereof in the same manner.
4. Recompute features and similarities for each application.
5. Retrain classifier to obtain λ' and ρ' .
6. Compute gain to assess the effectiveness of each obfuscation.

Case Study on Extended-Brennan-Greenstadt Corpus

- 45 authors
- 6500 words per author minimum
- We divided texts of each author into 3 artificial communities, resulting in 135 identities.
- We analyzed **33 different feature classes** and used a **linear SVM approach**.

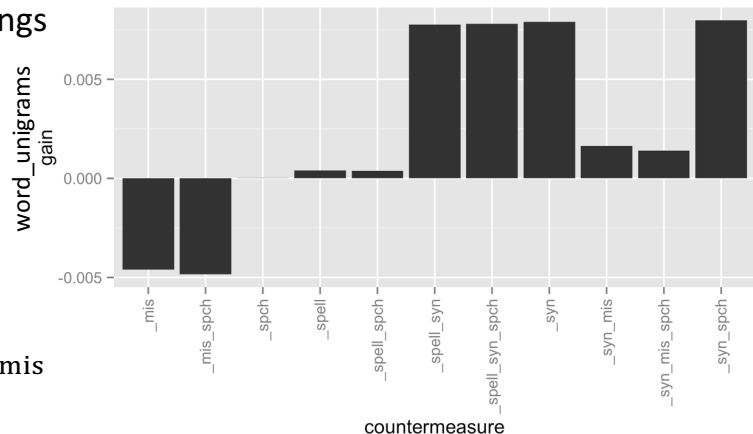
Automated Obfuscation Results

- We analyzed the 4 obfuscation techniques and some of their combinations.
- Except for one case, a **positive gain** also implied a drop in the **accuracy**.
- Only some features have been touched by our techniques/countermeasures.



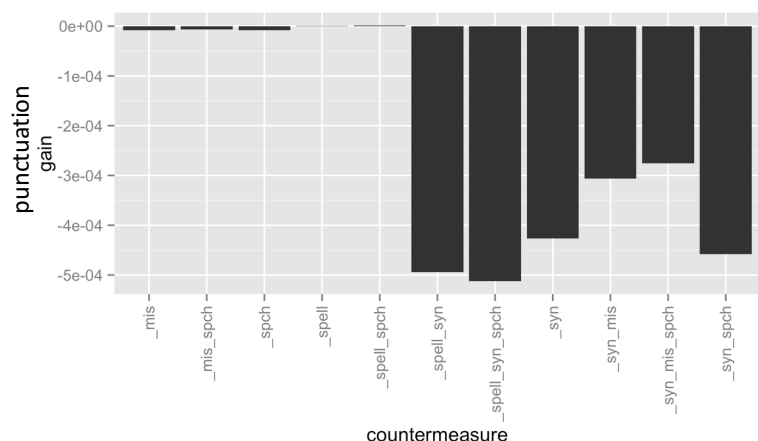
Results – Expected

- Most changes happen for the *word unigram* feature class.
- Some propagate to the *word bigrams*.
- While adding misspellings sometimes helps to get more similar to another identity, we might be still too far away for a positive gain.
- In our case study, the misspellings technique added too much, resulting in better identifiability.
- Gains seem to be more or less additive, e.g.,
 $\text{gain}^{\text{syn}+\text{mis}} \approx \text{gain}^{\text{syn}} + \text{gain}^{\text{mis}}$



Results – Interesting

- All obfuscations that involve our synonym replacement tool have a negative gain in the *punctuation* feature class.
- In our case study, the synonym replacement tool often changed single words to hyphenated compound words, e.g., *bad* → *high-risk*.
- This, later on, makes those identities better identifiable w.r.t. punctuation.



Results

- Top 4 obfuscation techniques ordered by their overall gain:
 1. Spellchecking + Synonym Replacement
 2. Spellchecking + Synonym Replacement + Special Characters
 3. Synonym Replacement
 4. Synonym Replacement + Special Characters
- After that, we have a larger drop in the gain.
- **Lessons learned:** fully automated obfuscation is not easy,
- **But:** we can formally evaluate its effect and gain useful insights.

Overview

General Linkability Model

- Linkability of Online Profiles
- d-convergence

Measuring Anonymity using d-convergence

- Anonymity Estimation
- Experimental Evaluation

Authorship Attribution as a Linkability Problem

- Intro to Stylometry
- Model for Countermeasure Effectiveness
- Experimental Evaluation