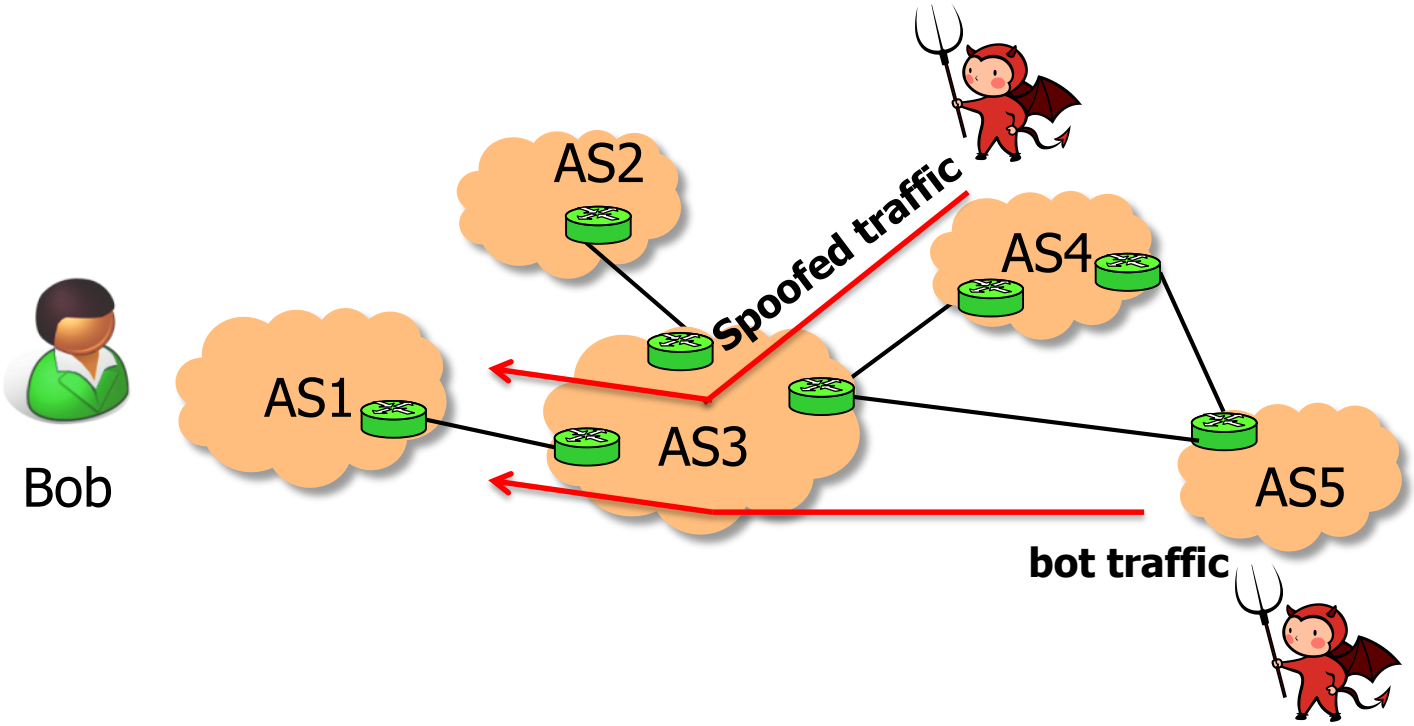
PRISM: Private Retrieval of the Internet's Sensitive Metadata

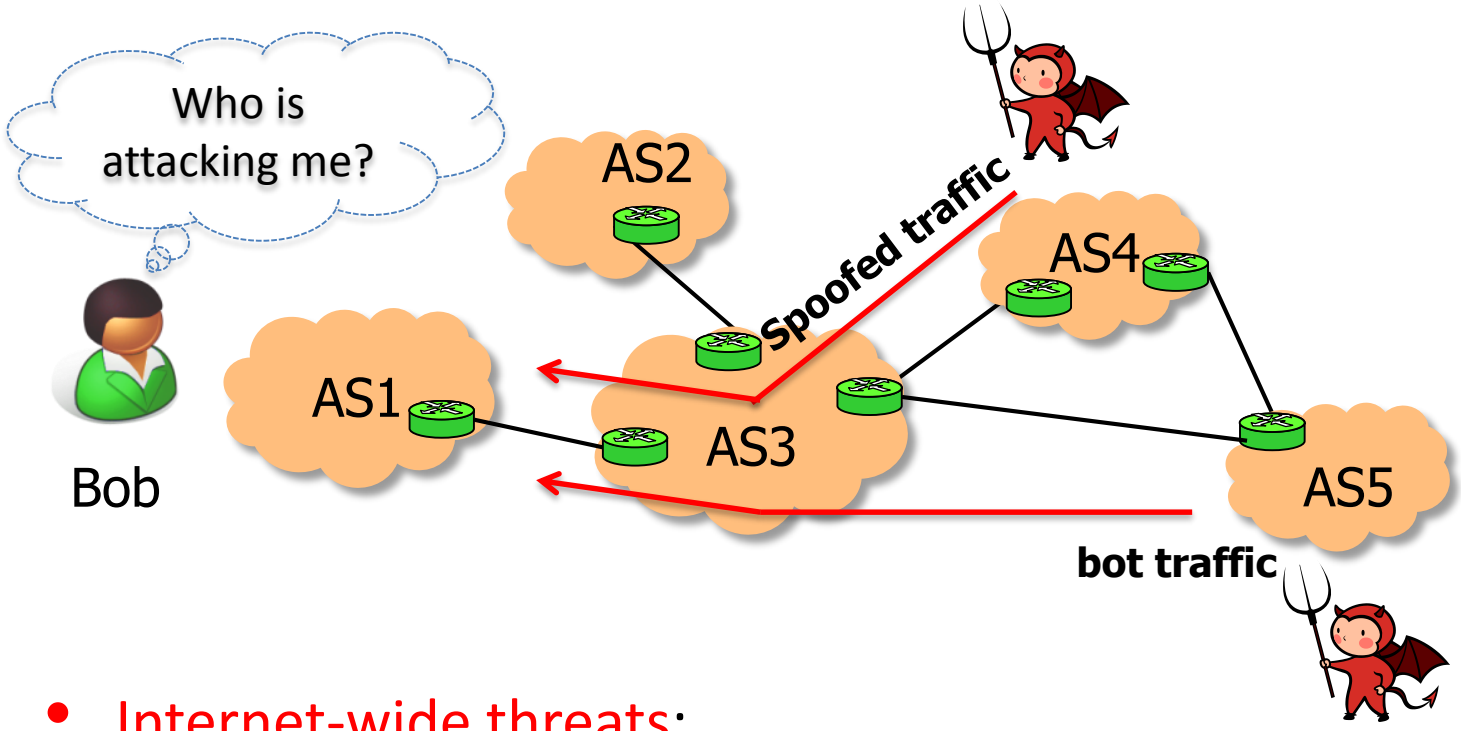
Ang Chen **Andreas Haeberlen**

University of Pennsylvania

Motivation: Internet-wide threats



Motivation: Internet-wide threats

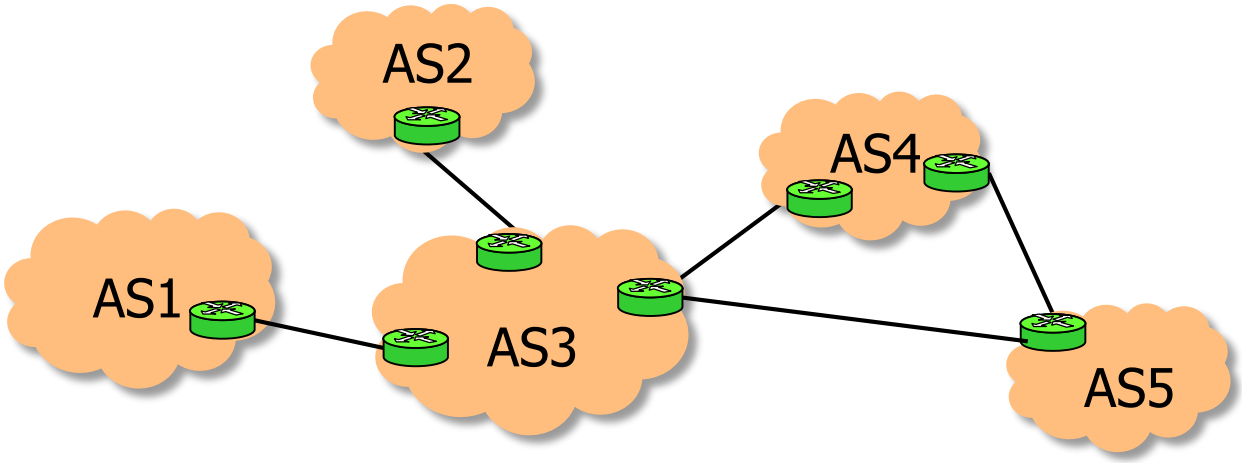


- Internet-wide threats:
  - Example: Botnet detection, DDoS backtrace, …
  - Bots scattered in many domains
  - But victims only see local ‘views’.

Having multiple data sources helps

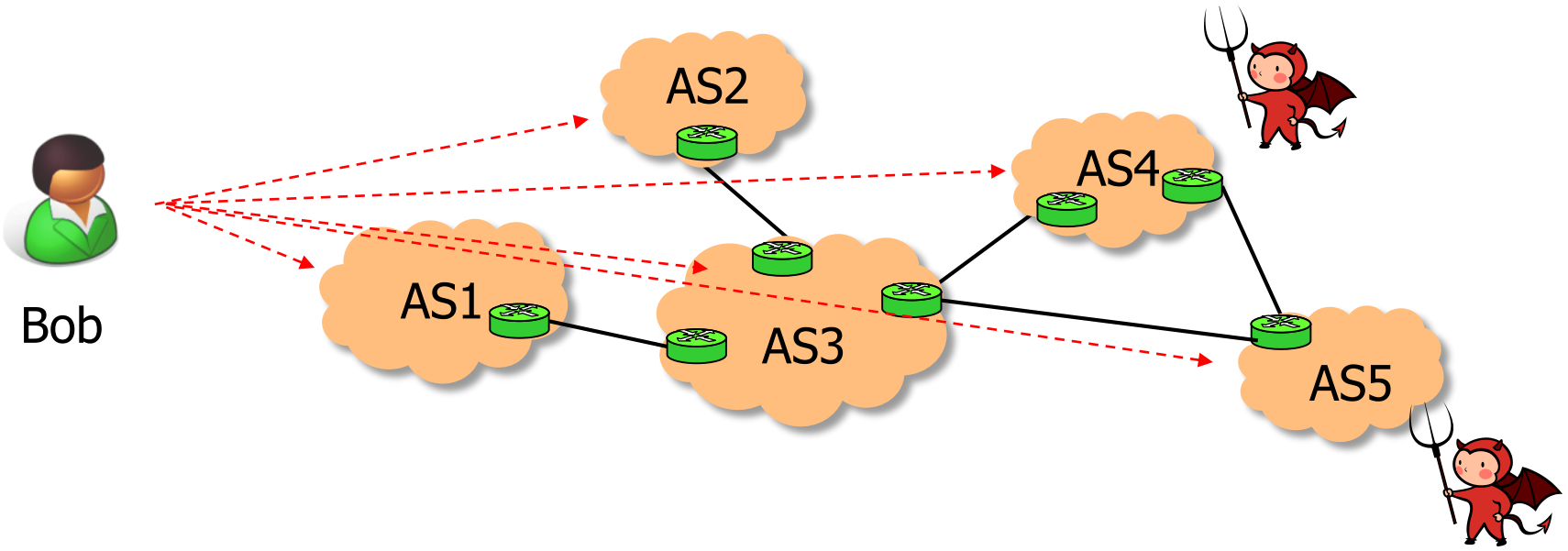


Bob



- Detect attacks using multiple domains' data
  - Multiple data sources are better than one!

Having multiple data sources helps

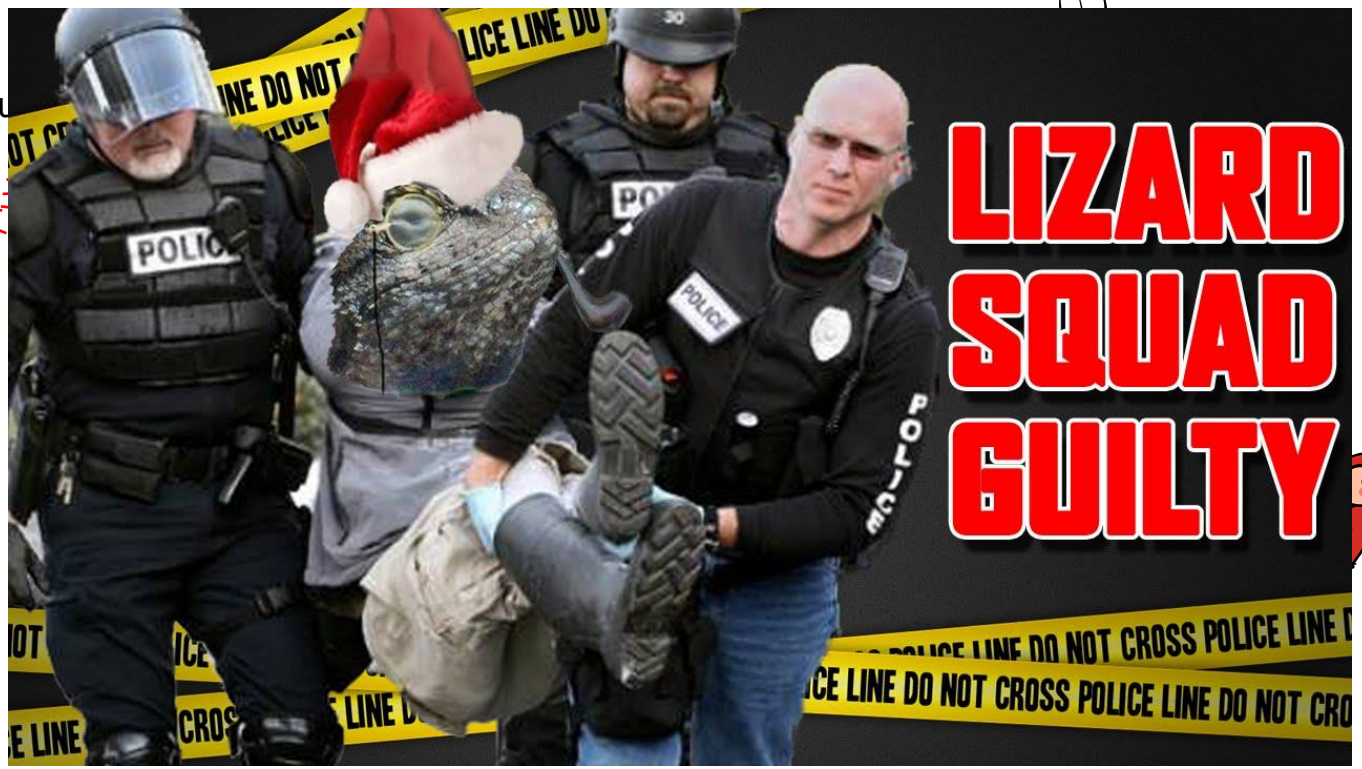


- Detect attacks using multiple domains' data
  - Multiple data sources are better than one!

Having multiple data sources helps



Bob



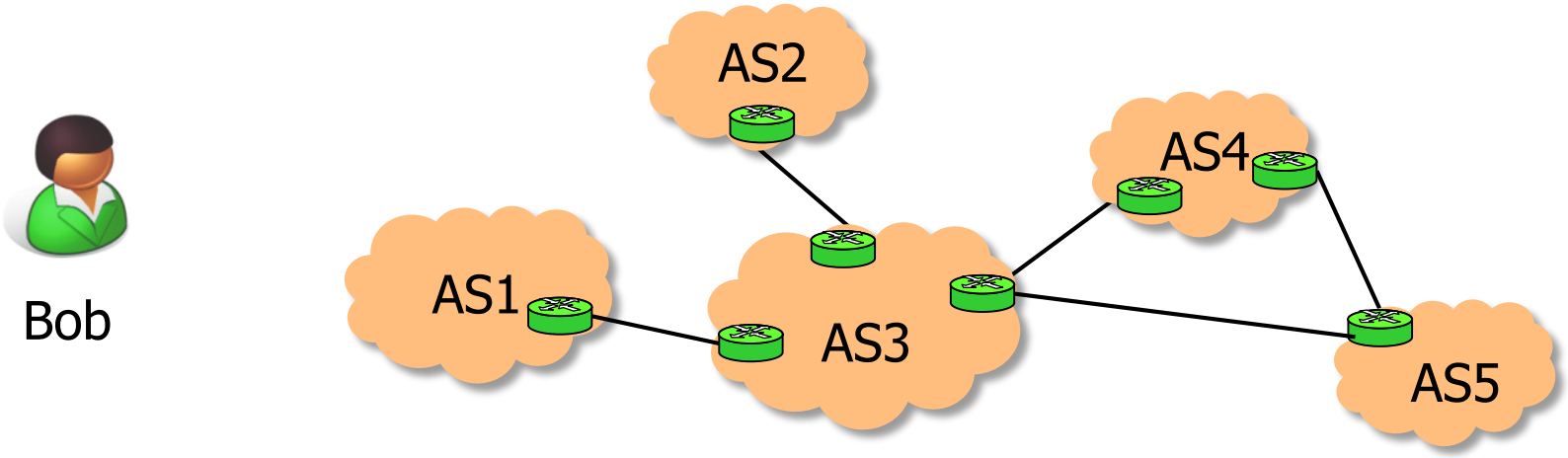
- Detect attacks using multiple domains' data
  - Multiple data sources are better than one!

Having multiple data sources helps



- Detect attacks using multiple domains' data
  - Multiple data sources are better than one!
  - Example: DDoS detection with **98%** accuracy on four domains' data [Chen-TPDS-2007]

Simple to write, hard to implement



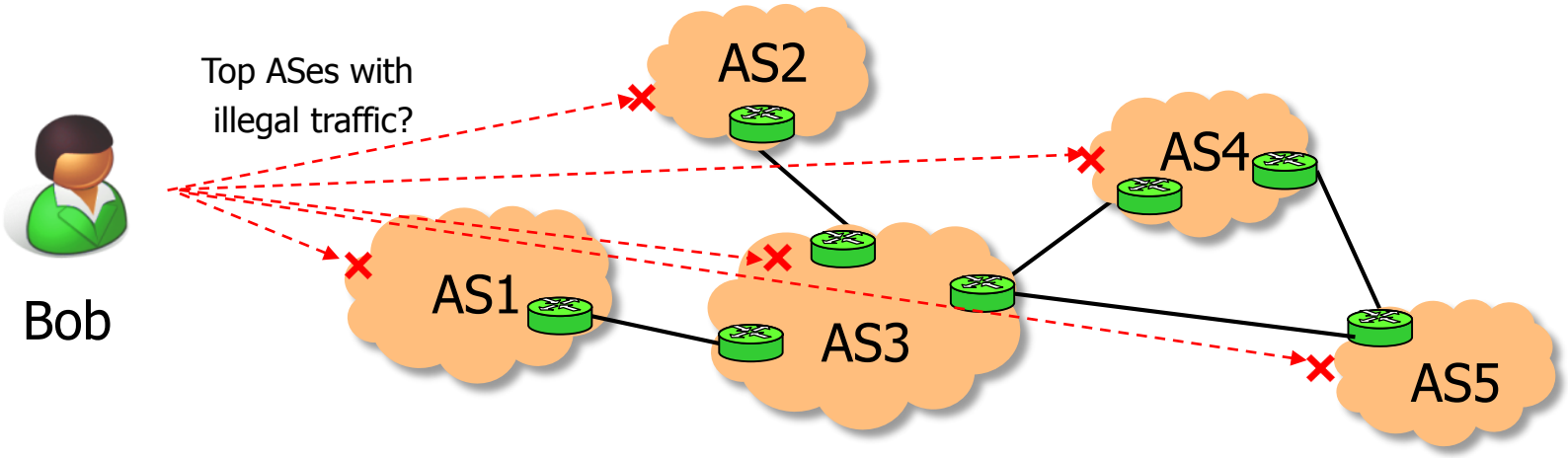
- Toy example: top ASes that generate darknet traffic:

  SELECT TOP 10 flow.SourceAS

  FROM JOIN Internet **BY** FlowID

  WHERE flow.destIP **IN** Darknet

Simple to write, hard to implement



- Toy example: top ASes that generate darknet traffic:

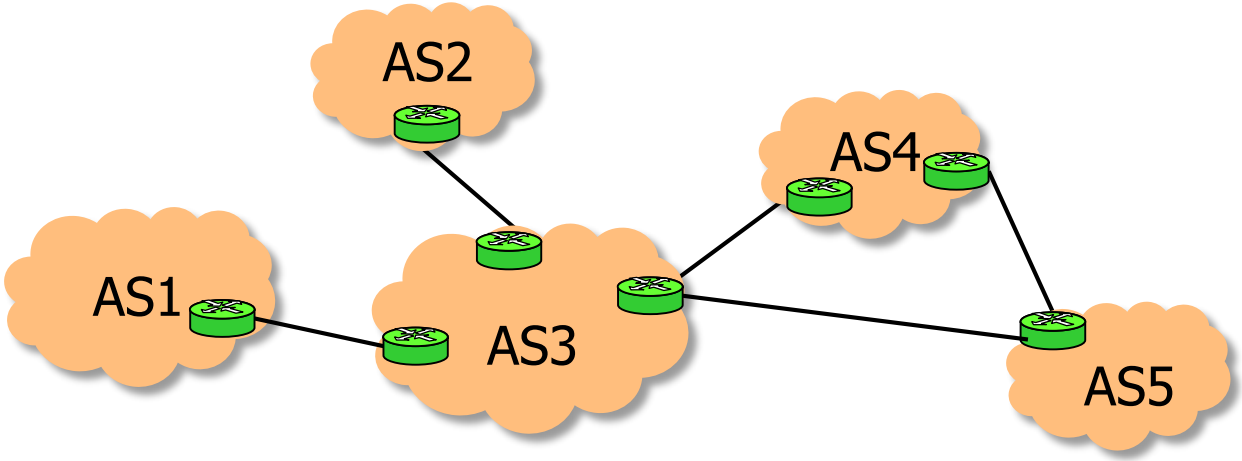
  SELECT TOP 10 flow.SourceAS

  FROM JOIN Internet **BY** FlowID

  WHERE flow.destIP **IN** Darknet

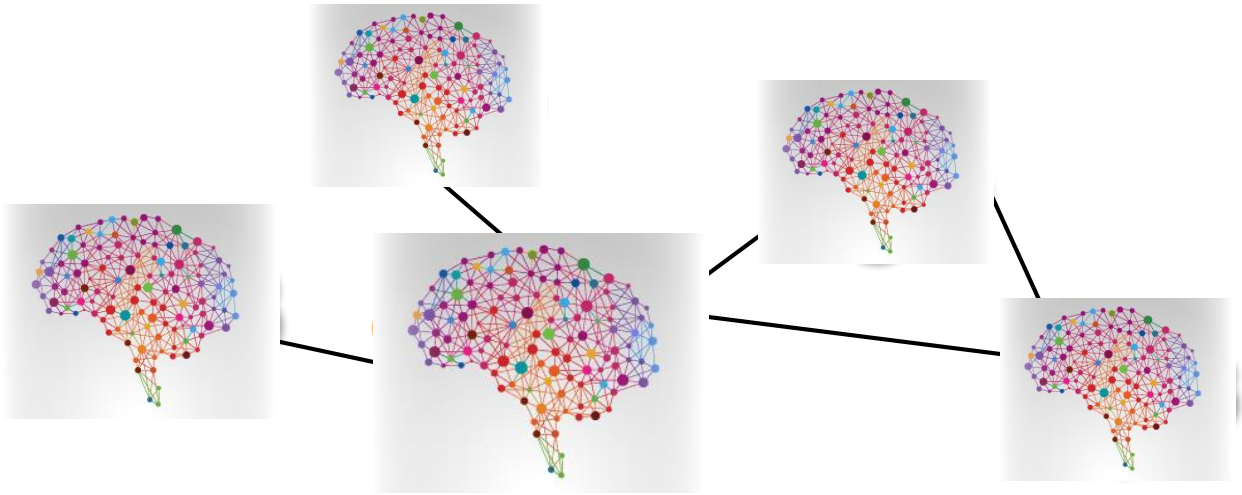
- Privacy concern: all data is not available in a single place!

An Internet "knowledge plane"



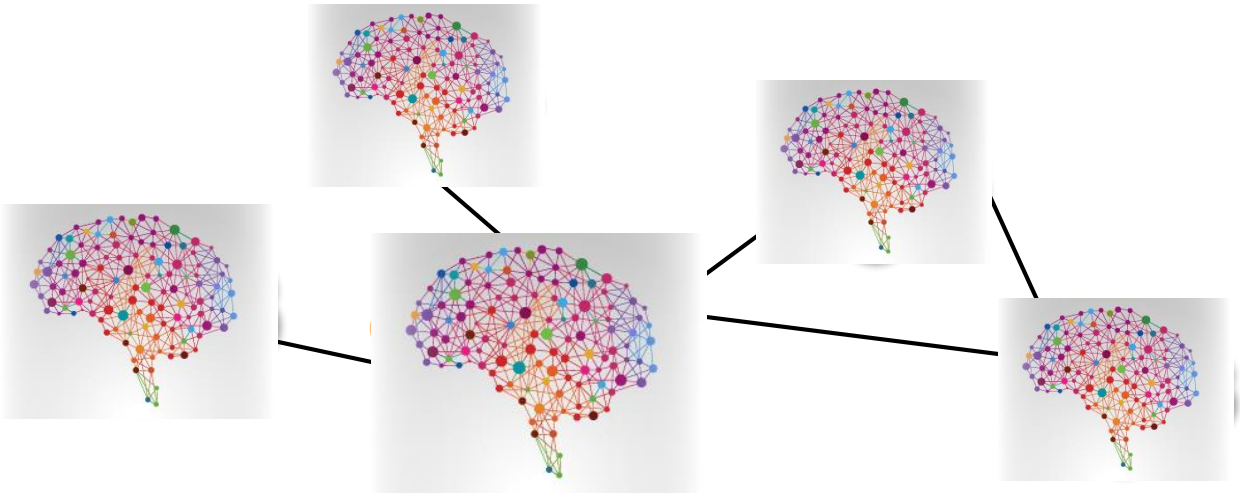
- A long-standing vision [Clark-SIGCOMM-2003]

An Internet "knowledge plane"



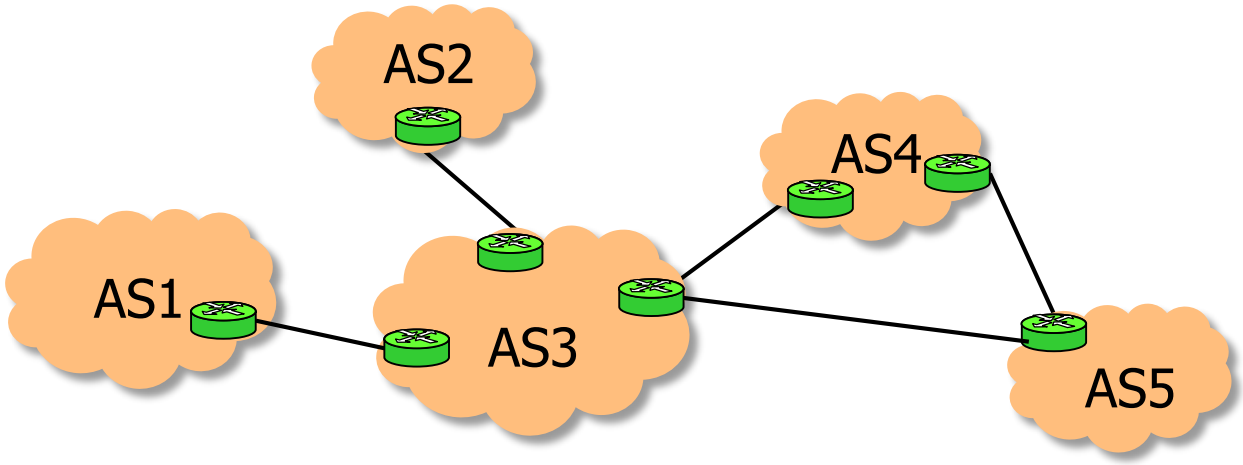
- A long-standing vision [Clark-SIGCOMM-2003]
  - Internet produces data about itself
  - Allow real-time queries on metadata
  - You can know what is happening where, when

An Internet "knowledge plane"

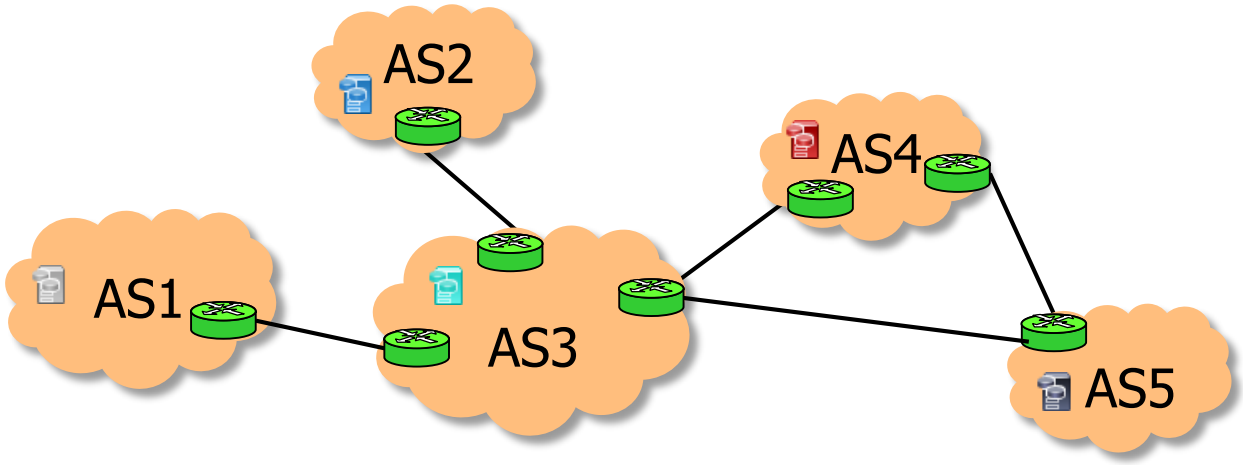


- A long-standing vision [Clark-SIGCOMM-2003]
  - Internet produces data about itself
  - Allow real-time queries on metadata
  - You can know what is happening where, when
- Benefits:
  - DDoS backtrace, botnet analysis, distributed troubleshooting, distributed forcasting…

What does it take to make this work?



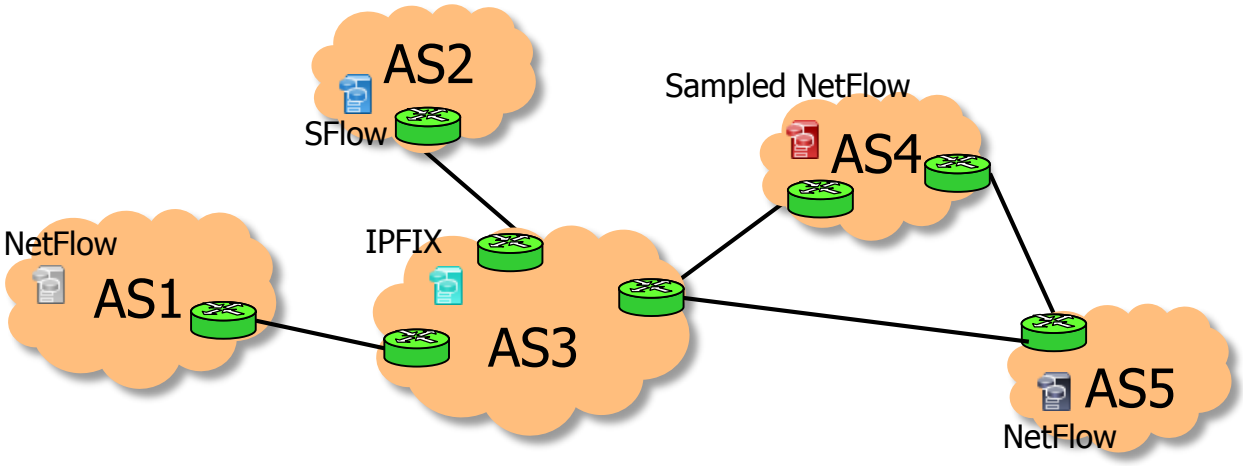
What does it take to make this work?





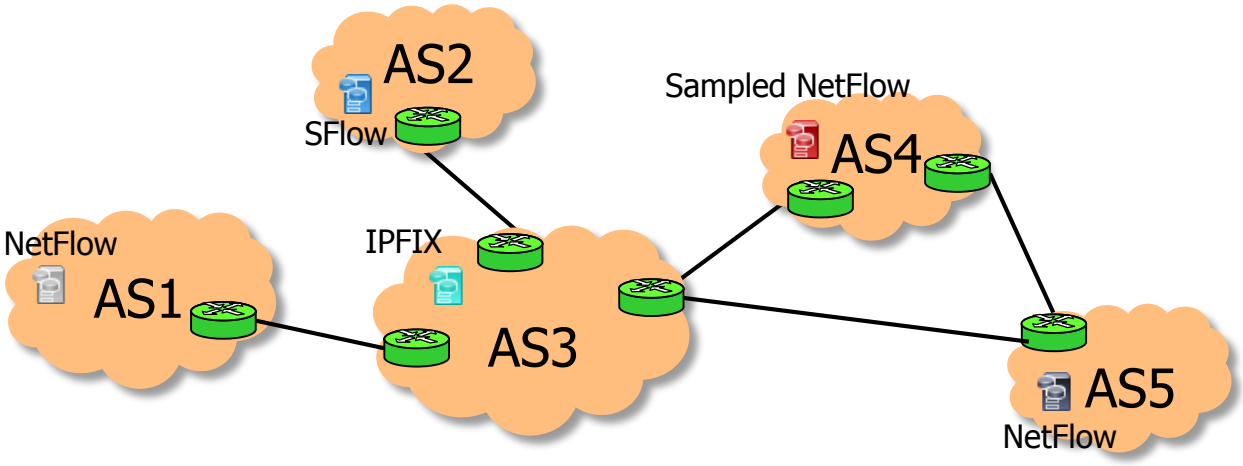
- Domains produce data about their operations.

What does it take to make this work?



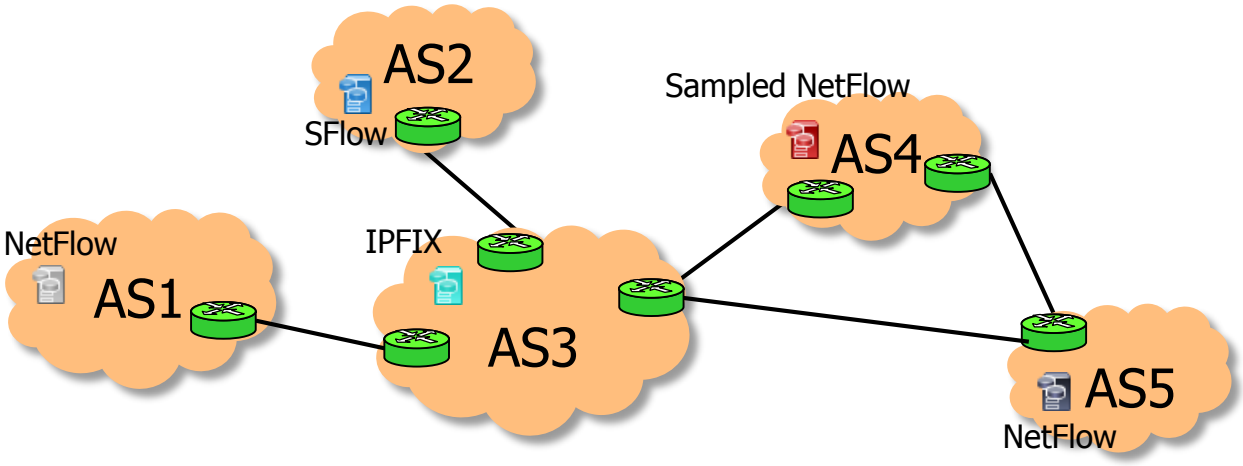
- Domains produce data about their operations.
- Domains use similar data formats.

What does it take to make this work?



- Domains produce data about their operations.
- Domains use similar data formats.
- Domains allow each other to query their data.

What does it take to make this work?



- Domains produce data about their operations.
- Domains use similar data formats.
- Domains allow each other to query their data.

Why are domains reluctant to share data?



Netflix de-anonymization



AOL searcher exposed

- Privacy is difficult even if you have the best intentions
  - Even after anonymization (Netflix de-anonymization case)

Why are domains reluctant to share data?



Netflix de-anonymization



AOL searcher exposed

- Privacy is difficult even if you have the best intentions
  - Even after anonymization (Netflix de-anonymization case)
  - Or aggregation (auxiliary information attack)

Why are domains reluctant to share data?



Netflix de-anonymization



AOL searcher exposed

- Privacy is difficult even if you have the best intentions
  - Even after anonymization (Netflix de-anonymization case)
  - Or aggregation (auxiliary information attack)
- To make a ‘knowledge plane’ work, we need strong privacy guarantees!

Why are domains reluctant to share data?



Netflix de-anonymization



AOL searcher exposed

- Privacy is difficult even if you have the best intentions
  - Even after anonymization (Netflix de-anonymization case)
  - Or aggregation (auxiliary information attack)
- To make a 'knowledge plane' work, we need strong privacy guarantees!
  - Idea: differential privacy.

Differential privacy

- Differential privacy:
  - What: provide very strict privacy guarantee for individuals.
  - 'Worst-case' adversary
  - Tunable amount of privacy
  - Composable query costs

Differential privacy

- Differential privacy:
  - What: provide very strict privacy guarantee for individuals.
  - 'Worst-case' adversary
  - Tunable amount of privacy
  - Composable query costs
- But, there are caveats too:
  - Limited query budget.
  - Gives noised answer.
  - Distributed DP is hard.
  - …

Differential privacy

- Differential privacy:
  - What: provide very strict privacy guarantee for individuals.
  - 'Worst-case' adversary
  - Tunable amount of privacy
  - Composable query costs
- But, there are caveats too:
  - Limited query budget.
  - Gives noised answer.
  - Distributed DP is hard.
  - …

Differential privacy: a good candidate?



Differential privacy

- Differential privacy:
  - What: provide very strict privacy guarantee for individuals.
  - 'Worst-case' adversary
  - Tunable amount of privacy
  - Composable query costs
- But, there are caveats too:
  - Limited query budget.
  - Gives noised answer.
  - Distributed DP is hard.
  - …

Differential privacy: a good candidate?

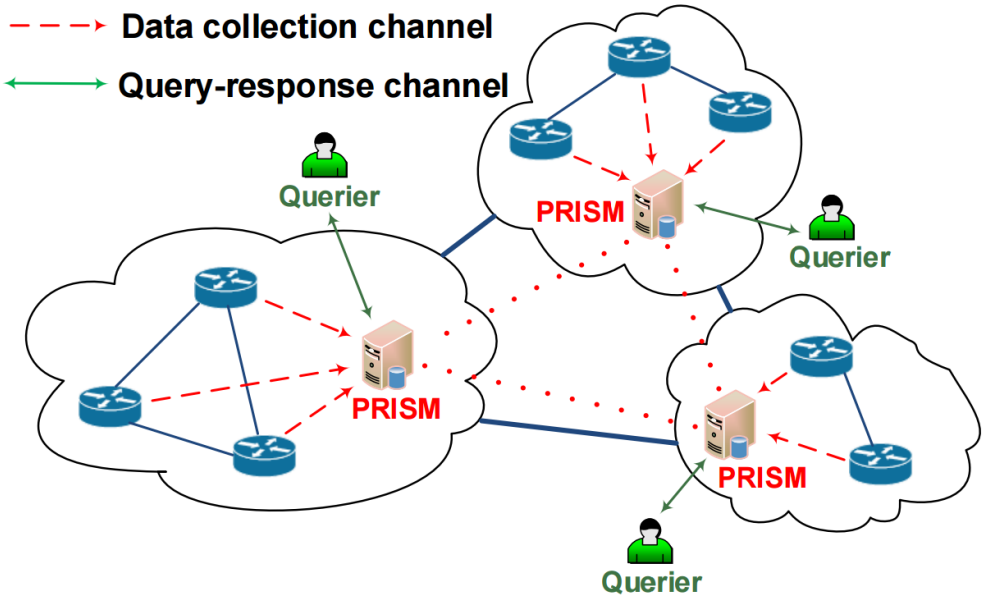


Our hypothesis: Yes!

Outline

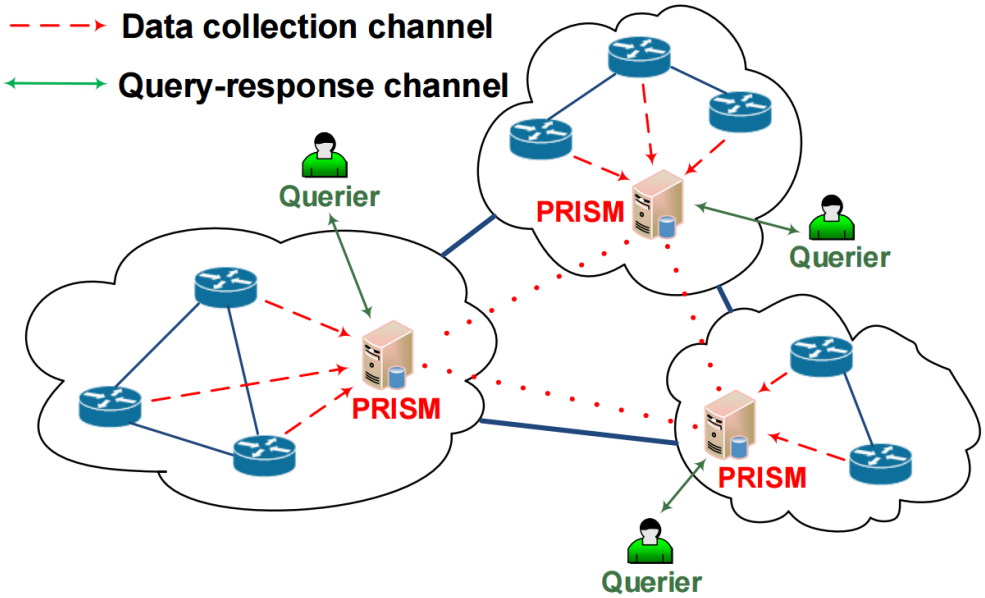
- ✓ - Motivation
- ✓ - Challenges
- - PRISM: Private Retrieval of the Internet's Sensitive Metadata
  - - The vision
  - - Do we have enough budget?
  - - What about data quality?
  - - Can we deal with attackers?
  - - Can we answer all types of queries?
  - - What about privacy for ISPs?
- - Conclusion

PRISM: differential privacy on Internet data



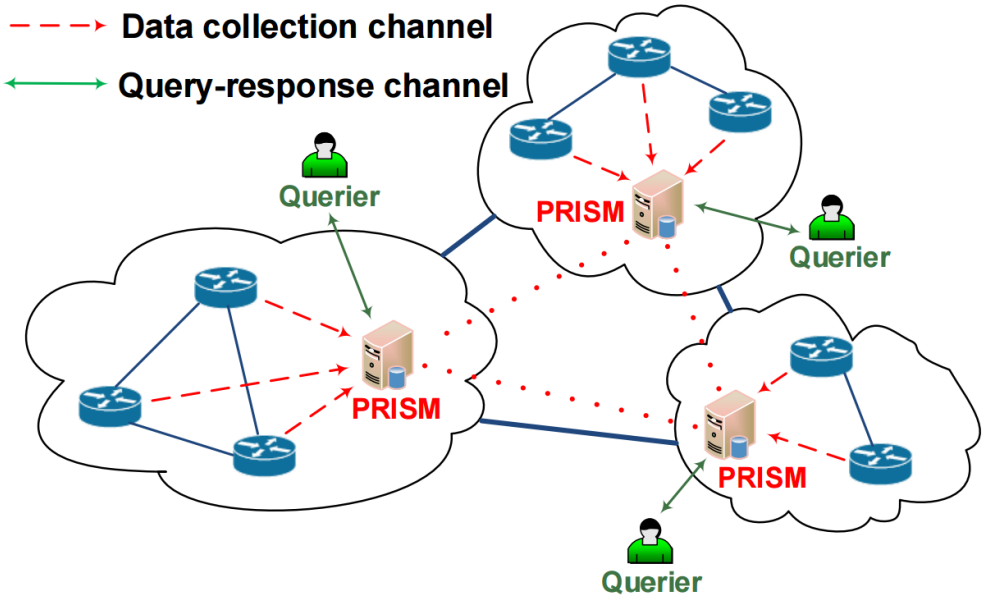
- PRISM: a system sketch

PRISM: differential privacy on Internet data



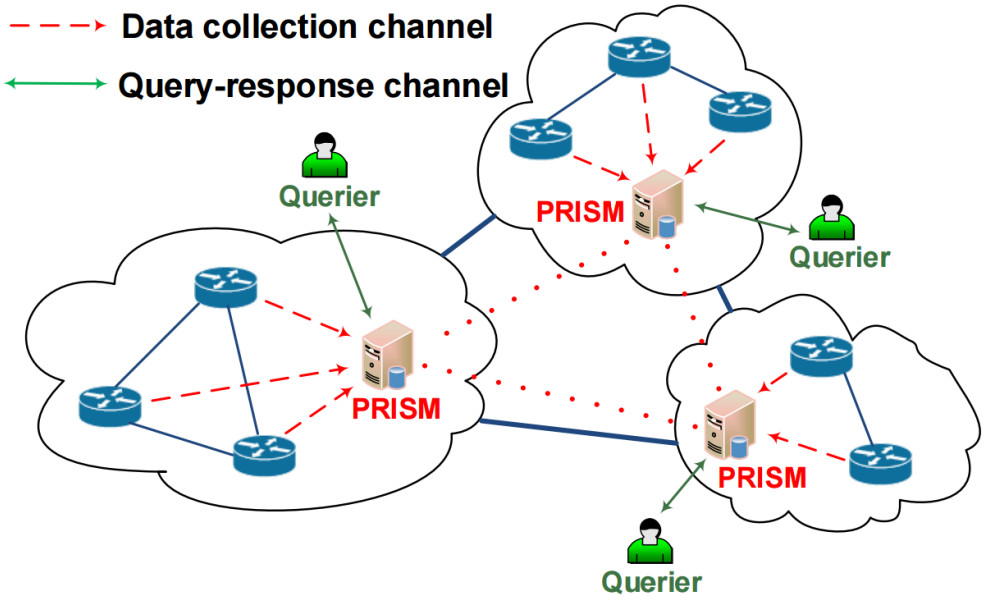
- PRISM: a system sketch
  - Domains keep their data local.

PRISM: differential privacy on Internet data



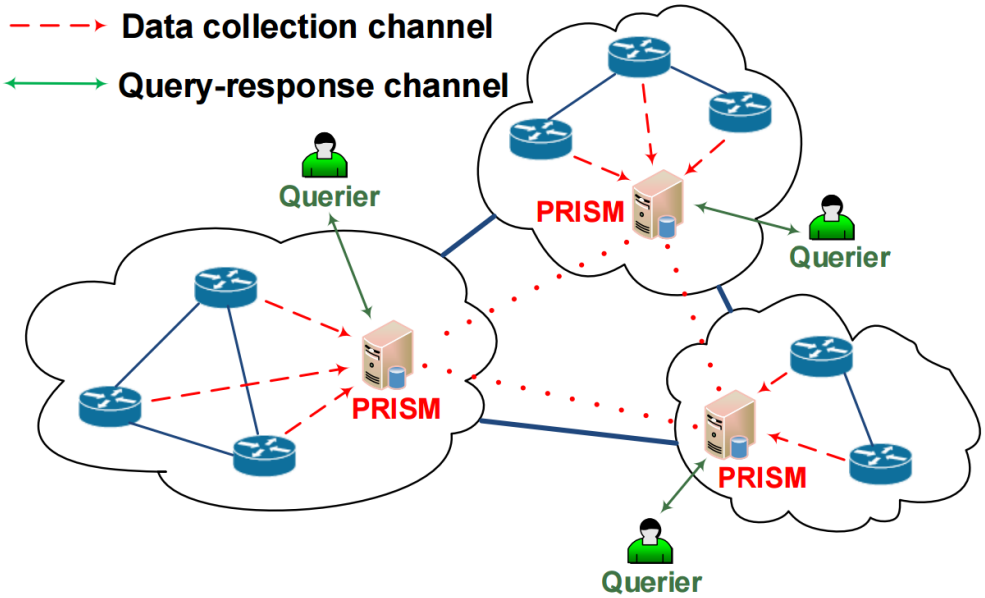
- PRISM: a system sketch
  - Domains keep their data local.
  - PRISM nodes manage local data and answer queries.

PRISM: differential privacy on Internet data



- PRISM: a system sketch
  - Domains keep their data local.
  - PRISM nodes manage local data and answer queries.
  - Query answers released with differential privacy.

PRISM: differential privacy on Internet data



- PRISM: a system sketch
  - Domains keep their data local.
  - PRISM nodes manage local data and answer queries.
  - Query answers released with differential privacy.
- Result: private Internet knowledge plane

Background: Differential privacy

- How: noise query answer before release
  - E.g., noise drawn from a Laplace distribution parameterized by ε.
  - ε: privacy parameter; larger values = more privacy release.

Background: Differential privacy

- How: noise query answer before release
  - E.g., noise drawn from a Laplace distribution parameterized by ε.
  - ε: privacy parameter; larger values = more privacy release.
- Guarantee:
  - Query answer on ‘neighboring databases’ are very similar.

Background: Differential privacy

- How: noise query answer before release
  - E.g., noise drawn from a Laplace distribution parameterized by ε.
  - ε: privacy parameter; larger values = more privacy release.
- Guarantee:
  - Query answer on 'neighboring databases' are very similar.
- We can view ε as a privacy budget:
  - The total amount of privacy we are willing to release.
  - Each query uses up some budget.
  - Refuse further queries once budget is depleted.

Challenges

- Do we have enough budget?
- Can we detect attacks with noised data?
- What about compromised PRISM nodes?
- Does PRISM provide privacy for ISPs, too?
- Would PRISM work with a partial deployment?
- Can we make all queries differentially private?
- Would PRISM's query processor scale?
- …

Challenges

- Do we have enough budget?
- Can we detect attacks with noised data?
- What about compromised PRISM nodes?
- Does PRISM provide privacy for ISPs, too?
- Would PRISM work with a partial deployment?
- Can we make all queries differentially private?
- Would PRISM's query processor scale?
- …

Challenges

- Do we have enough budget?
- Can we detect attacks with noised data?
- What about compromised PRISM nodes?
- Does PRISM provide privacy for ISPs, too?
- Would PRISM work with a partial deployment?
- Can we make all queries differentially private?
- Would PRISM's query processor scale?
- …

See paper

The privacy budget

- Admin can set their own privacy budget ε.
- Differential privacy is composable:
  - Two queries with budget ε1 and ε2 costs the same with one query with budget (ε1+ε2).
  - PRISM continues answering queries until ε runs out.
  - Estimation of number of queries: noised answer is within ±E of the true answer with probability c.

$$N = \frac{\varepsilon \cdot E}{-2 \cdot s \cdot \ln(1-c)}$$

- The budget problem: ε sets a hard limit on how many queries PRISM can answer.
  - Many ways to set ε [e.g., Hsu-CSF-2013]
  - No matter how large, budget eventually runs out.

Challenge #1: enough budget?

- The Internet data presents unique opportunities!

Challenge #1: enough budget?

- The Internet data presents unique opportunities!
- Large size: queries cost less.

Challenge #1: enough budget?

- The Internet data presents unique opportunities!

- Large size: queries cost less.
  - E.g., counting queries about IP addresses.
  - Assume that the answer is 40 million, we want released answer to be 10% within true answer with 95% confidence

Challenge #1: enough budget?

- The Internet data presents unique opportunities!

- Large size: queries cost less.
  - E.g., counting queries about IP addresses.
  - Assume that the answer is 40 million, we want released answer to be 10% within true answer with 95% confidence
  - N = 667,616.

Challenge #1: enough budget?

- The Internet data presents unique opportunities!

- Large size: queries cost less.
  - E.g., counting queries about IP addresses.
  - Assume that the answer is 40 million, we want released answer to be 10% within true answer with 95% confidence
  - N = 667,616.
  - Per ISP: ~10 queries.

Challenge #1: enough budget?

- Sampling: reduces query cost

Challenge #1: enough budget?

- Sampling: reduces query cost
  - Internet data is typically sampled, e.g., NetFlow is typically sampled at 1/4K.

Challenge #1: enough budget?

- Sampling: reduces query cost
  - Internet data is typically sampled, e.g., NetFlow is typically sampled at 1/4K.
  - Theoretical result: sampling at rate α reduces cost to $\alpha * \varepsilon$.

Challenge #1: enough budget?

- Sampling: reduces query cost
  - Internet data is typically sampled, e.g., NetFlow is typically sampled at 1/4K.
  - Theoretical result: sampling at rate α reduces cost to $\alpha * \varepsilon$.
  - We further sample NetFlow records by ~50%.
  - Per ISP: ~100,000 queries.

Challenge #1: enough budget?

- We probably don’t have a worst-case adversary!

Challenge #1: enough budget?

- We probably don't have a worst-case adversary!
  - ISPs are competitors, so won't collude on a large scale.

Challenge #1: enough budget?

- We probably don't have a worst-case adversary!
  - ISPs are competitors, so won't collude on a large scale.
  - Conservatively, if no two ISPs collude, we can give each ISP its own budget.
  - This scales up budget significantly.

Challenge #1: enough budget?

- We probably don't have a worst-case adversary!
  - ISPs are competitors, so won't collude on a large scale.
  - Conservatively, if no two ISPs collude, we can give each ISP its own budget.
  - This scales up budget significantly.
  - Even there are small-scale collusions, per ISP: 400 million queries are within reach (1K queries per ISP per day for 1,000 years.)

Challenge #1: enough budget?

- Can we replenish the budget?

Challenge #1: enough budget?

- Can we replenish the budget?
  - Internet data is fast changing
    - E.g., many flows expire in seconds
    - E.g., IP-to-user mappings also change
    - E.g., 40% of /24 address blocks are dynamic

Challenge #1: enough budget?

- Can we replenish the budget?
  - Internet data is fast changing
    - E.g., many flows expire in seconds
    - E.g., IP-to-user mappings also change
    - E.g., 40% of /24 address blocks are dynamic
  - Eventually, the DB may become entirely different, e.g., in 100 years, most users should be different.

Challenge #1: enough budget?

- Can we replenish the budget?
  - Internet data is fast changing
    - E.g., many flows expire in seconds
    - E.g., IP-to-user mappings also change
    - E.g., 40% of /24 address blocks are dynamic
  - Eventually, the DB may become entirely different, e.g., in 100 years, most users should be different.
  - There should be opportunity for replenishing the budget when users are completely different.

Challenge #2: data quality?

- The data quality problem: if DP adds noise, can we still detect attacks accurately?

Challenge #2: data quality?

- The data quality problem: if DP adds noise, can we still detect attacks accurately?

- DP's noise is easy to interpret!
  - Well-known distribution: Laplace.
  - Dealing with imprecision: well understood topic.

Challenge #2: data quality?

- The data quality problem: if DP adds noise, can we still detect attacks accurately?

- DP's noise is easy to interpret!
  - Well-known distribution: Laplace.
  - Dealing with imprecision: well understood topic.
  - Works on true data: instead of inferred data.

Challenge #2: data quality?

- The data quality problem: if DP adds noise, can we still detect attacks accurately?

- DP's noise is easy to interpret!
  - Well-known distribution: Laplace.
  - Dealing with imprecision: well understood topic.
  - Works on true data: instead of inferred data.
  - We are looking for large trends, e.g., DDoS, bots.

Challenge #3: compromised nodes?

- What if PRISM nodes are compromised?

Challenge #3: compromised nodes?

- What if PRISM nodes are compromised?
- There are things we can do, too!

Challenge #3: compromised nodes?

- What if PRISM nodes are compromised?
- There are things we can do, too!
  - Hackers are unlikely to take over the majority of nodes.

Challenge #3: compromised nodes?

- What if PRISM nodes are compromised?

- There are things we can do, too!
  - Hackers are unlikely to take over the majority of nodes.
  - Quality-checking can be integrated with queries. [Reed-2010-ICFP]

Challenge #3: compromised nodes?

- What if PRISM nodes are compromised?

- There are things we can do, too!
  - Hackers are unlikely to take over the majority of nodes.
  - Quality-checking can be integrated with queries. [Reed-2010-ICFP]
  - Queries answers can be released verifiably [Narayan-2015-Eurosys]

Other challenges

- Challenge #4: Difficult queries
- Challenge #5: Privacy for ISPs
- Challenge #6: Partial deployment
- Challenge #7: Scaling the query processor
- …

Other challenges

- Challenge #4: Difficult queries
- Challenge #5: Privacy for ISPs
- Challenge #6: Partial deployment
- Challenge #7: Scaling the query processor
- …

Please read paper for details.

Conclusion

- Motivation: Internet-wide threats
- Primary challenge: privacy concern
- Proposal: PRISM
  - Differential privacy for Internet data
- Feasibility
  - Privacy budget
  - Noised data for detection?
  - Compromised nodes?
  - …

Conclusion

- Motivation: Internet-wide threats
- Primary challenge: privacy concern
- Proposal: PRISM
  - Differential privacy for Internet data
- Feasibility
  - Privacy budget
  - Noised data for detection?
  - Compromised nodes?
  - …

Questions?