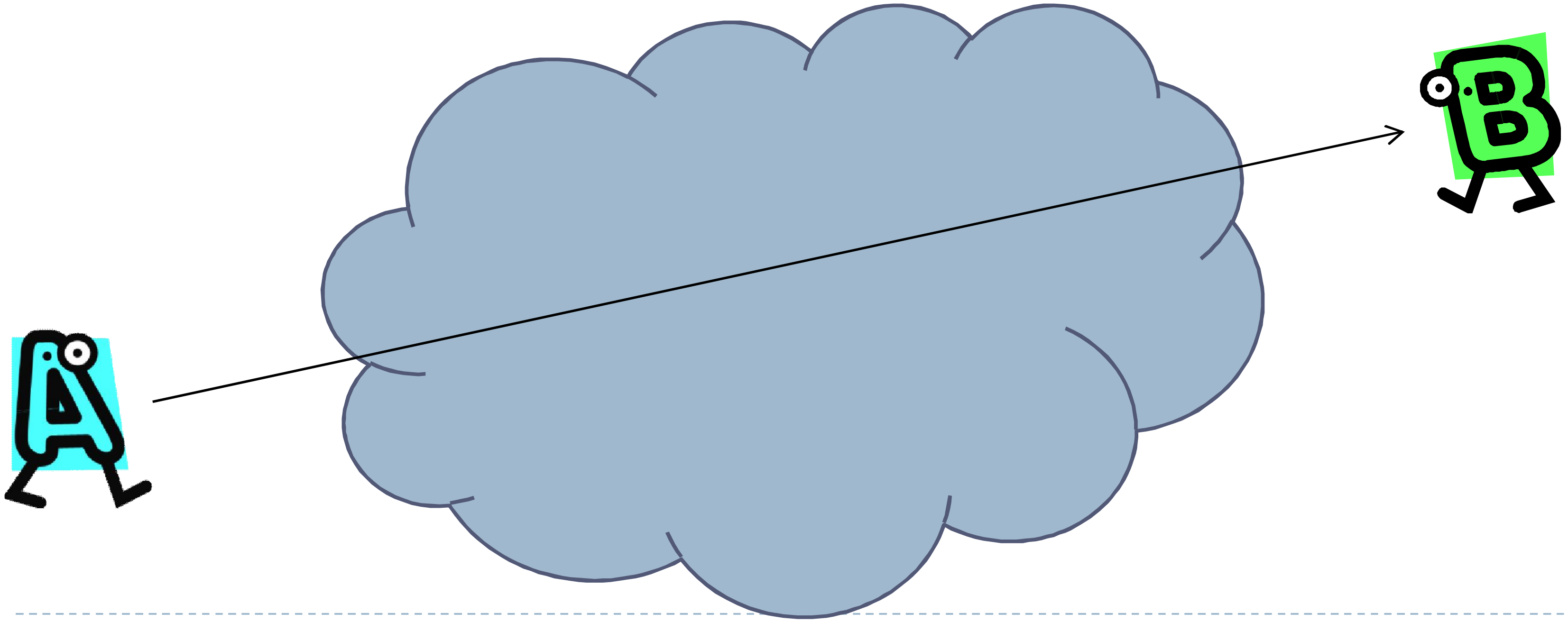
The Bayesian Traffic Analysis of Mix Networks

Carmela Troncoso
George Danezis

CCS -- 11 November 2009
Microsoft Research Cambridge / KU Leuven(COSIC)

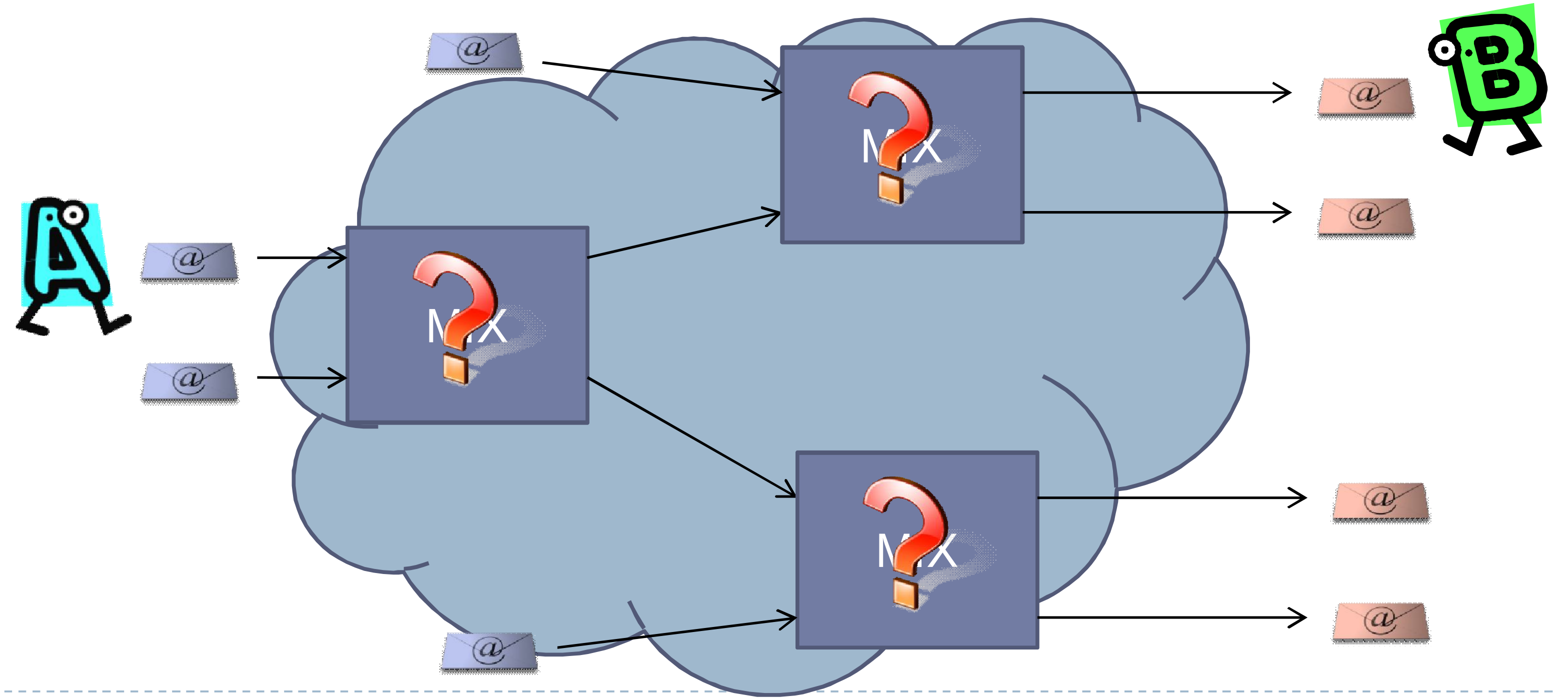
Anonymity

- Motivation
  - “Tell me who your friends are…”
  - Election protocols (e-voting)
  - Freedom of speech



Mix networks

- Mixes hide relations between inputs and outputs
- Mixes are combined in networks in order to
  - Distribute trust (one good mix is enough)
  - Load balancing (no mix is big enough)



Attacks against mix networks

- Uncover who speaks to whom
  - Observe all links (*Global Passive Adversary)*

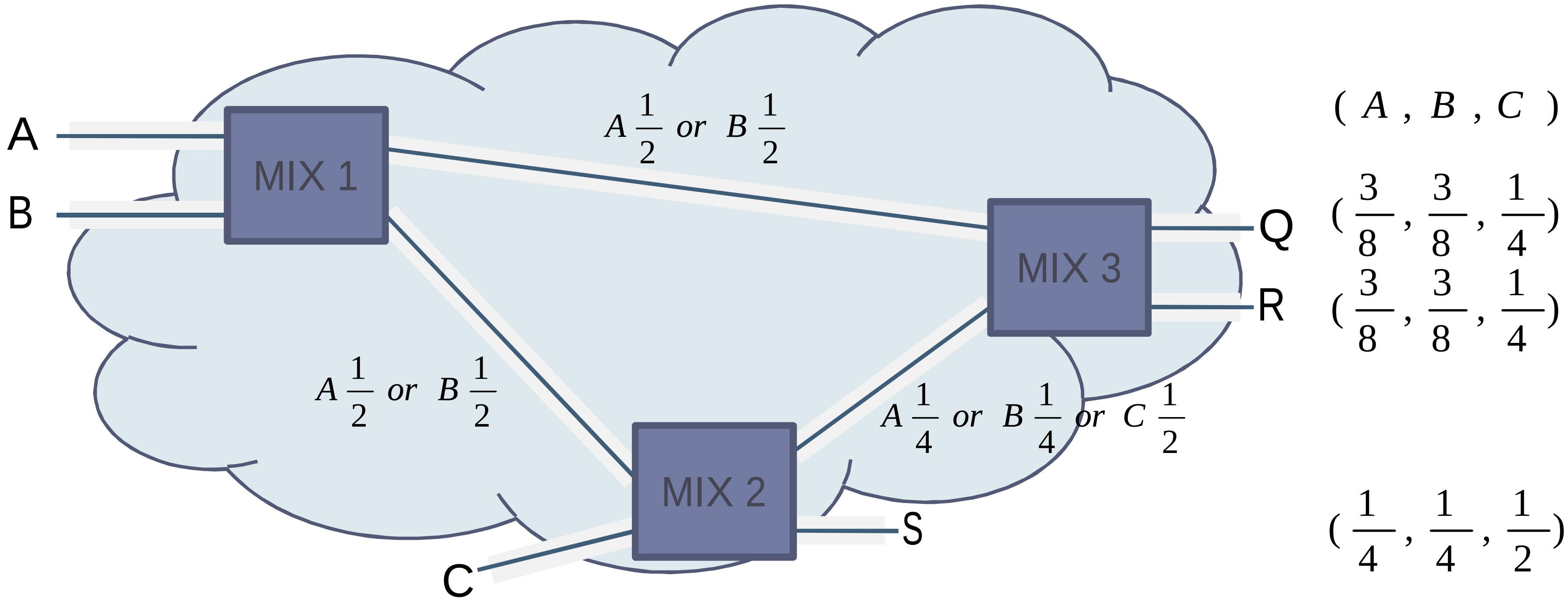
  - Restricted routes [Dan03]
    - Messages cannot follow any route
  - Bridging and Fingerprinting [DanSyv08]
    - Users have partial knowledge of the network

  - Long term disclosure attacks:
    - Exploit persistent patterns
    - Disclosure Attack [Kes03], Statistical Disclosure Attack [Dan03], Perfect Matching Disclosure Attacks [Tron-et-al08]

- Based on heuristics and specific models, not generic

Mix networks and traffic analysis

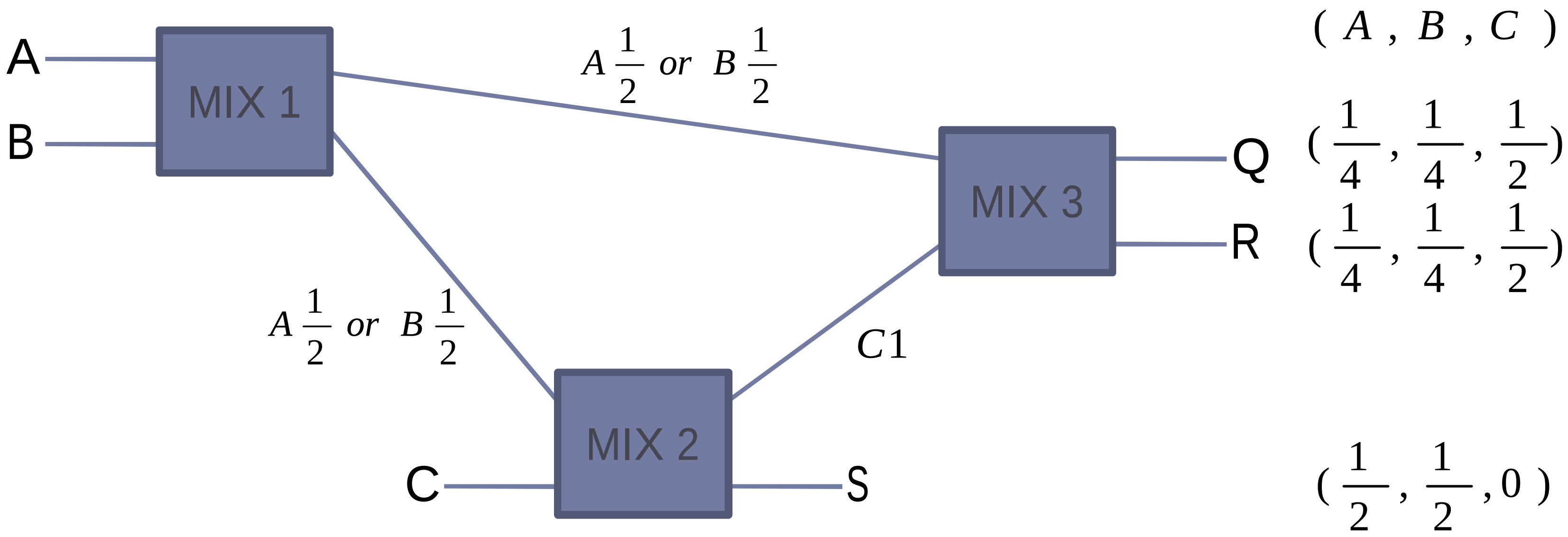
- Determine probability distributions input-output



- Threshold mix: collect t messages, and outputs them changing their appearance and in a random order

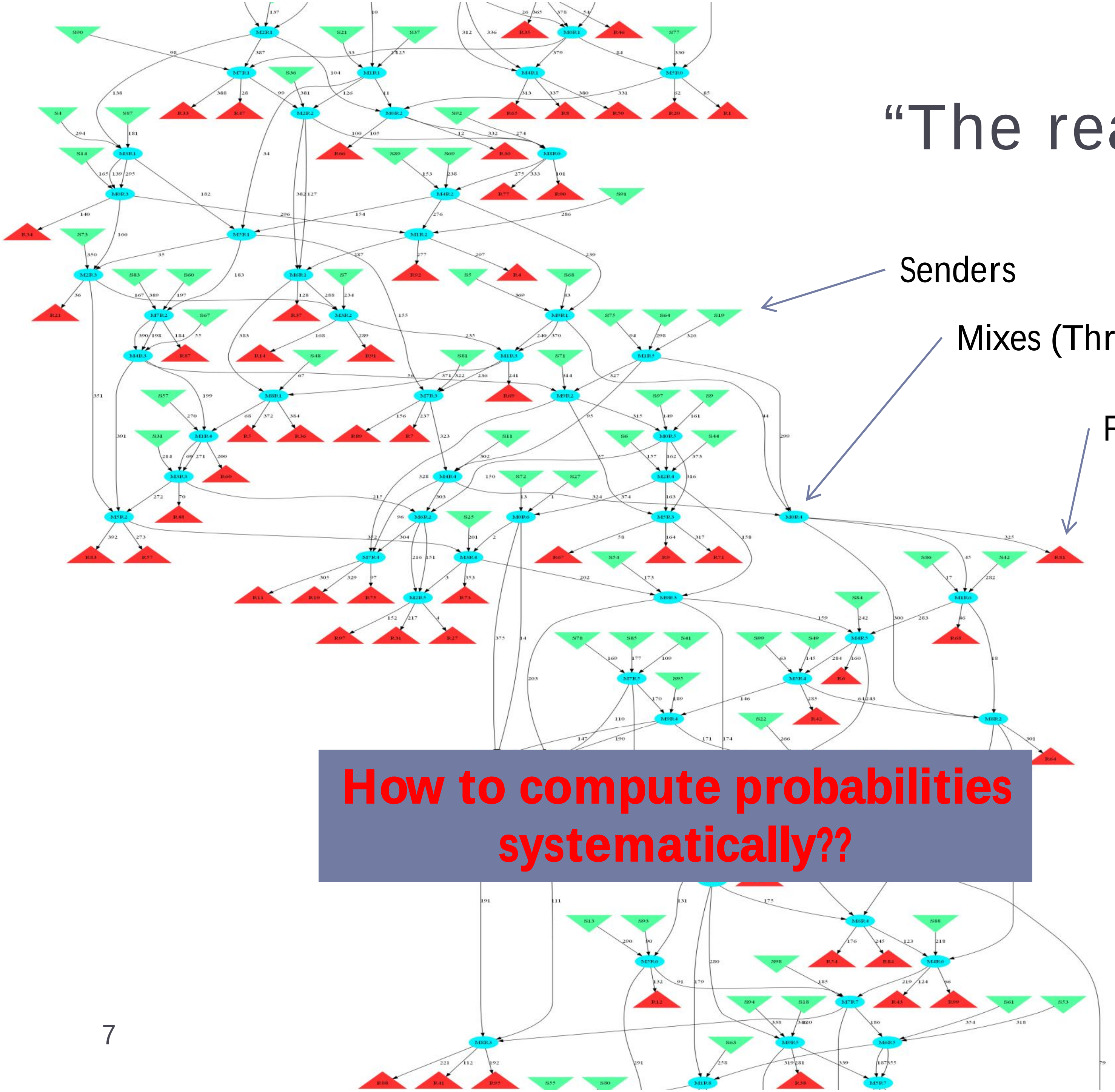
Mix networks and traffic analysis

- Constraints, e.g. length=2



Non trivial given observation!!

“The real thing”



Senders

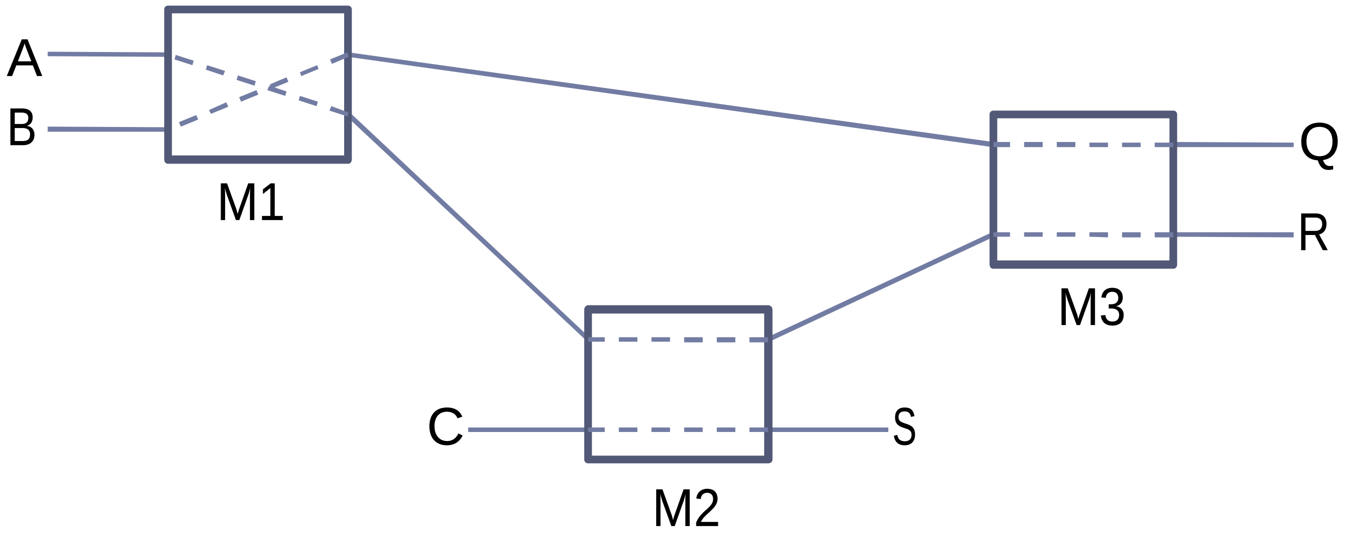
Mixes (Threshold = 3)

Receivers

How to compute probabilities systematically??

Redefining the traffic analysis problem

- *"Hidden State" + Observation = Paths*



$P_1 \rightarrow$ A —— M1 —— M2 —— M3 —— R

$P_2 \rightarrow$ B —— M1 —— M3 —— Q

$P_3 \rightarrow$ C —— M2 —— S

$$\Pr(HS \mid O, C) = \frac{\Pr(O \mid HS, C) \cdot \Pr(HS \mid C)}{\sum_{HS} \Pr(HS, O \mid C)} = \frac{\Pr(Paths \mid C)}{\sum_{Paths} \Pr(Paths \mid C)}$$

Probabilistic model of mix networks

- Users decide their Paths independently

$$\Pr(Paths \mid C) = \prod_x \Pr(P_x \mid C)$$

  - Length restrictions
  - Node choice restrictions, no repetitions

$$\Pr(P_x \mid C) = \Pr(L = l \mid C) \cdot \Pr(M_x \mid L = l, C) \cdot I_{set}(M_x)$$

- Non-compliant clients (with probability $p_{\overline{cp}}$)
  - Do not respect length restrictions
  - Allow repetitions

$$\Pr(Paths \mid C) = \left[\prod_{i \in P_{\overline{cp}}} p_{\overline{cp}} \Pr(P_i \mid C, I_{\overline{cp}}(P_i)) \right] \cdot \left[\prod_{j \in P_{cp}} (1 - p_{\overline{cp}}) \Pr(P_j \mid C) \right]$$

Sampling to estimate probabilities

- For real traces $\Pr(HS \mid O, C)$ is infeasible to compute analytically because there are too many Hidden States

$$\Pr(HS \mid O, C) = \frac{\Pr(O \mid HS, C) \cdot \Pr(HS \mid C)}{\sum_{HS} \Pr(HS, O \mid C)} = \frac{\Pr(Paths \mid C)}{Z}$$

- … but we only care about marginal distributions
  - Is Alice speaking to Bob? $\Pr(A \to B \mid O, C)$

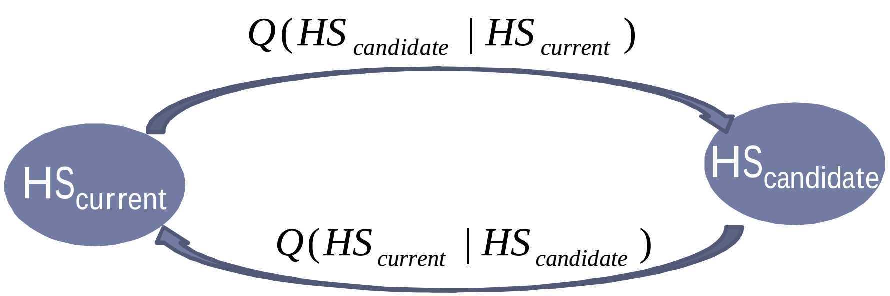
- We can calculate those if we have many samples of HS according to $\Pr(HS \mid O, C)$
  - We can simply count how many times Alice speaks to Bob

Markov Chain Monte Carlo

- Sample from a distribution difficult to sample from directly

$$\Pr(HS \mid O, C) = \frac{\Pr(O \mid HS, C) \cdot \Pr(HS \mid C)}{\sum_{HS} \Pr(HS, O \mid C)} = \frac{\Pr(O \mid HS, C) \cdot K}{Z} = \frac{\Pr(Paths \mid C)}{Z}$$

- Metropolis-Hastings sampling
  - Constructs a Markov Chain with stationary distribution $\Pr(HS \mid O, C)$
    - **Basic step:** Current state $\xrightarrow{Q}$ Candidate state



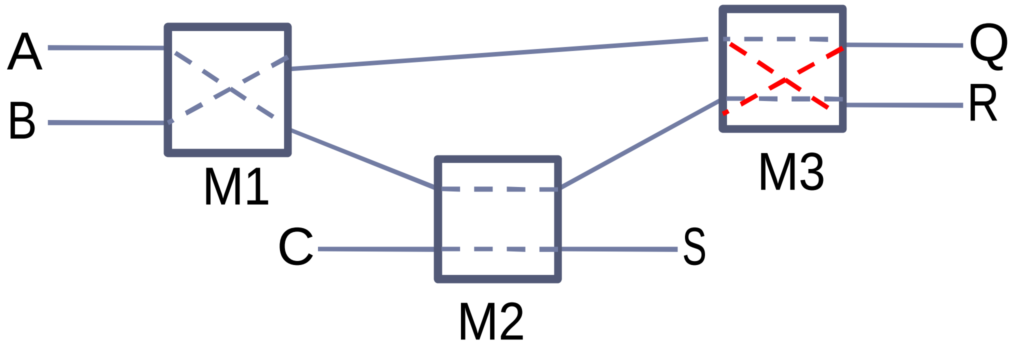
$$\alpha = \frac{\Pr(HS_{candidate})Q(HS_{candidate} \mid HS_{current})}{\Pr(HS_{current})Q(HS_{current} \mid HS_{candidate})}$$

$\alpha \geq 1$ **Go!**

$\alpha < 1$ **Go with probability α**

Our sampler: Q transition

- How do we propose candidate states?
- Transition Q: swap operation



  - More complicated transitions for non-compliant clients

- We get **independent samples of HS** by repeating this basic step many times before choosing a new sample

Evaluation

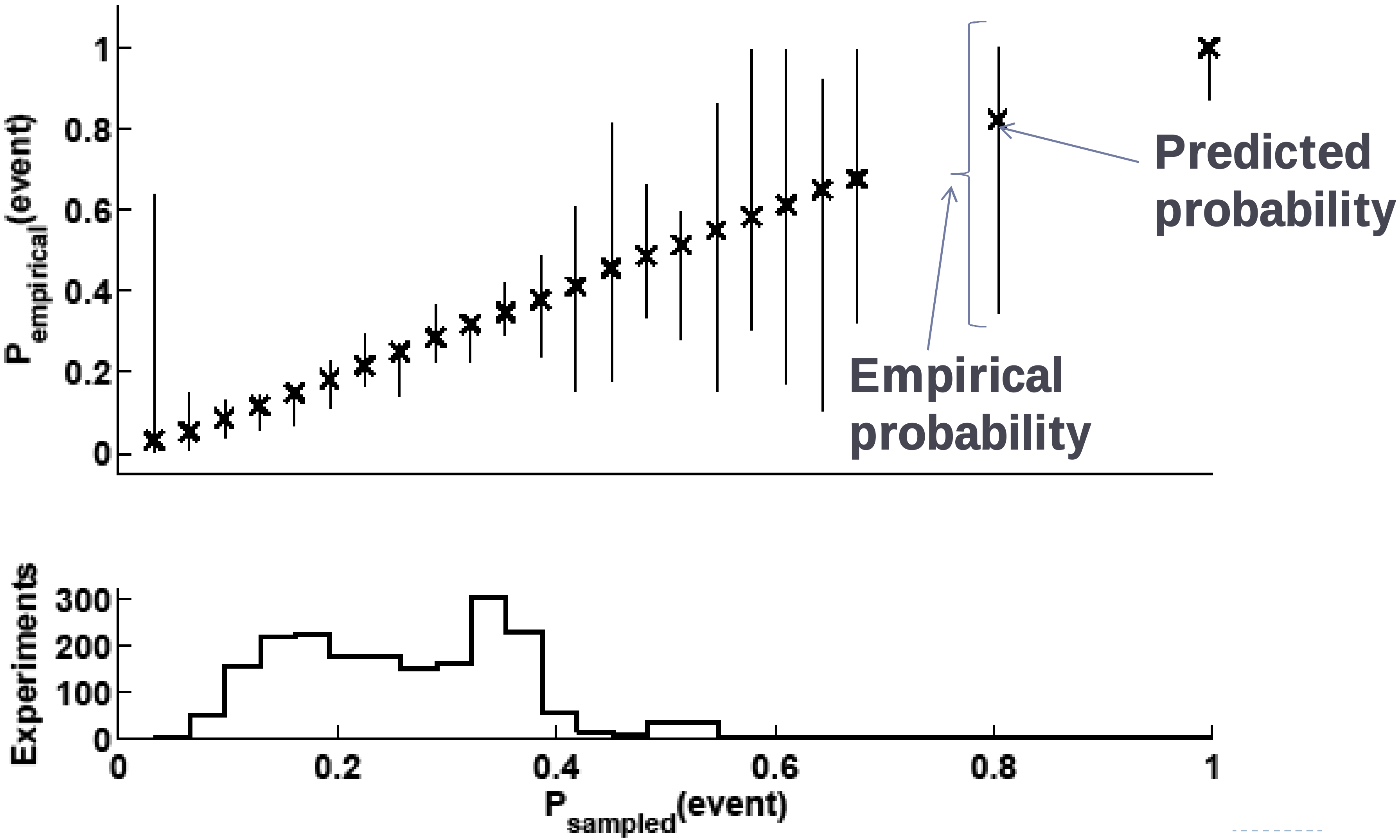
Events should happen with the predicted probability

1. Create an instance of a network
2. Run the sampler and obtain $P_1, P_2, \ldots$
3. Choose a target sender and a receiver
4. Predict probability

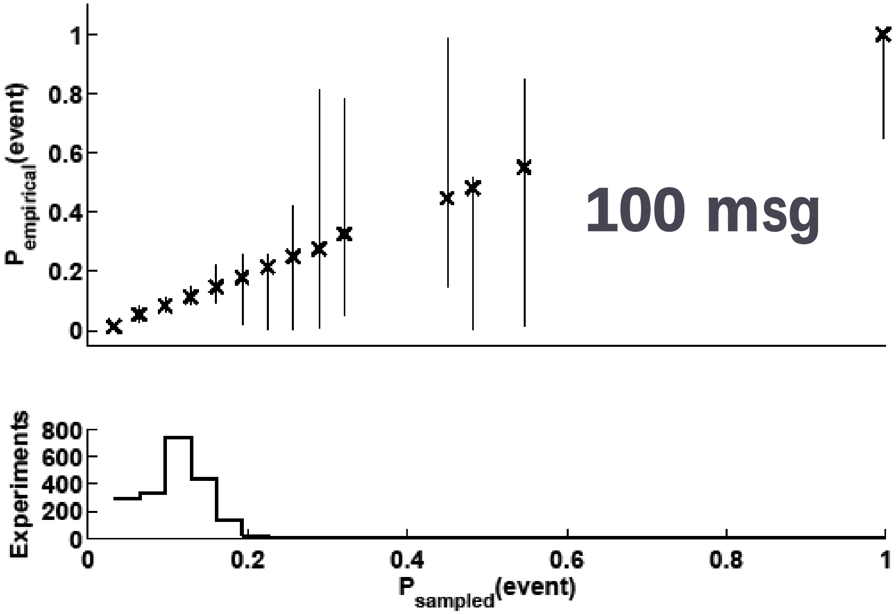
$$\Pr(\text{Sen} \rightarrow \text{Rec}) \approx \frac{\sum_j I_{\text{Sen}\rightarrow\text{Rec}}(Paths_j)}{N}$$

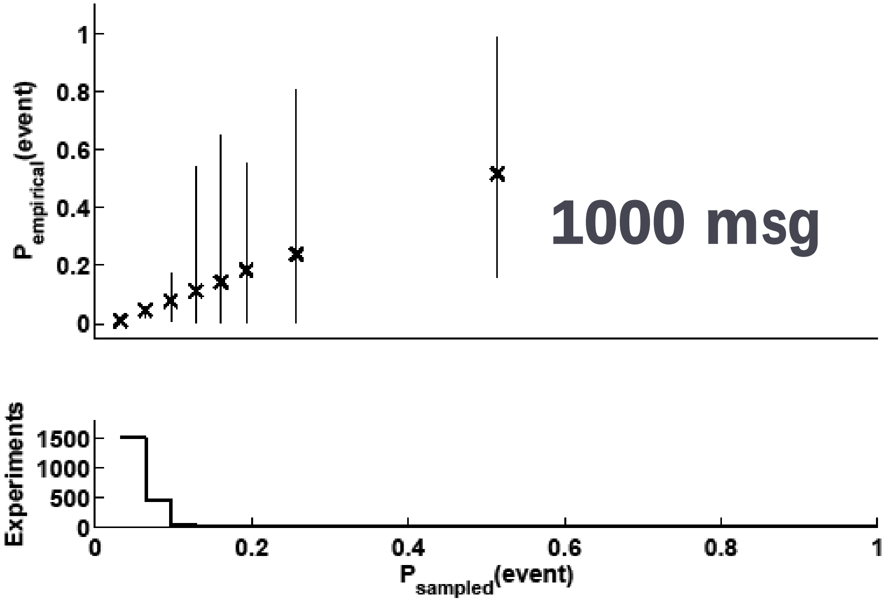
5. Check if actually Sen chose Rec as receiver $I_{\text{Sen}\rightarrow\text{Rec}}(network)$
6. Choose new network and go to 2

Results – 50 msg, compliant clients



Results – big networks





- It scales well as networks get larger
- As expected mix networks offer good protection

Performance

Nmix	t	Nmsg	RAM (Mb)	iter	Full analysis (min)	One sample (ms)
3	3	10	16	6011	4.24	509.12
3	3	50	18	6011	4.80	576.42
10	20	50	18	7011	5.34	641.28
10	20	1 000	24	7011	5.97	706.12
10	20	10 000	125	-	-	-

- RAM requirements
  - Size of network and population
- Time requirements (1443 LOC Python)
  - Operations are O(1)

Applications

- Evaluation information theoretic metrics for anonymity

$$H = -\sum_{R_i} P(A \to R_i \mid O, C) \cdot \log P(A \to R_i \mid O, C)$$

- Estimating probability of arbitrary events
  - Input message to output message?
  - Alice speaking to Bob ever?
  - Two messages having the same sender?

- Accommodate new constraints
  - Key to evaluate new mix network proposals

Conclusions

- Traffic analysis is non trivial when there are constraints
- Probabilistic model of mix networks: incorporates most attacks
  - Non-compliant clients
- Monte Carlo Markov Chain methods to extract marginal probabilities
- Key advantages:
  - Requires generative model (we know how to compute it!)
  - Systematically include all information available
  - Distribution over all possible states (not only most likely)

Time for questions

- If you liked this paper
  - **Vida: How to use Bayesian inference to de-anonymize persistent communications.** George Danezis and Carmela Troncoso. Privacy Enhancing Technologies Symposium 2009
  - **The Application of Bayesian Inference to Traffic analysis.** Carmela Troncoso and George Danezis Microsoft Technical Report
- If you want to see more similar research
  - **10th Privacy Enhancing Technologies (PETS)**
    - Berlin Jul 21 – Jul 23, 2010 - Deadline February 15
- ... if you miss the deadline and/or have some crazy idea you would like to discuss with the community
  - **HotPETS 2010** (deadline April 24)

http://petsymposium.org/