Bayesian Inference and Traffic Analysis

Carmela Troncoso
George Danezis

University of Texas at Arlington 16/11/2009
Microsoft Research Cambridge/ KU Leuven(COSIC)
Outline

- The traffic analysis problem and a bayesian approach
  - What is this all about

- Modelling mix networks
  - The boring part of the talk

- Markov Chain Monte Carlo methods and traffic analysis
  - The exciting part of the talk

- Evaluation and results
  - It actually works!

- Conclusions
Mix networks

- 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

\[ \frac{1}{2} A \text{ or } \frac{1}{2} B \]

\[ \frac{1}{4} A \text{ or } \frac{1}{4} B \text{ or } \frac{1}{2} C \]

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

\[ \left( \frac{3}{8}, \frac{3}{8}, \frac{1}{4} \right) \]

\[ \left( \frac{3}{8}, \frac{3}{8}, \frac{1}{4} \right) \]

\[ \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \right) \]
Mix networks and traffic analysis

- Constraints, e.g. length=2

Non trivial given observation!!

\[(A, B, C)\]

\[\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)\]

\[\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)\]

\[\left(\frac{1}{2}, \frac{1}{2}, 0\right)\]
"The real thing"

How to compute probabilities systematically??

Senders

Mixes (Threshold = 3)

 Receivers
Redefining the traffic analysis problem

- Find “hidden state” of the mixes

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

Too large to enumerate!!

Prior information
Redefining the traffic analysis problem

“Hidden State” + Observation = Paths

Pr($HS \mid O, C$) = $\frac{Pr(O \mid HS, C) \cdot K}{Z} = \frac{Pr(Paths \mid C)}{Z}$
Sampling to estimate probabilities (I)

- Actually... we want marginal probabilities

\[
\Pr( A \to Q \mid O, C) = \sum_{HS} \mathcal{I}_{A \to Q}(HS_j)
\]

- But... we cannot obtain them directly
Sampling to estimate probabilities (II)

- If we obtain samples
  \( HS_1, HS_2, HS_3, HS_4, \ldots, HS_N \sim \Pr( HS \mid O, C ) \)

  
  
  \[
  \text{Pr}(A \rightarrow Q \mid O, C) \approx \frac{\sum_{j} I_{A \rightarrow Q}(HS_j)}{N}
  \]

- Markov Chain Monte Carlo Methods

  \[
  \Pr( HS \mid O, C ) = \frac{\Pr( Paths \mid C )}{Z}
  \]

How does \( \Pr(Paths \mid C) \) look like?
Modelling mix networks
(The boring part of the talk)
Probabilistic model – Basic Constraints

- Users decide independently
  \[
  \Pr(Paths \mid C) = \prod_x \Pr(P_x \mid C)
  \]

- Length restrictions \( \Pr(L = l \mid C) \) with any distribution
  \[
  \text{e.g. uniform } (L_{\text{min}}, L_{\text{max}}) \quad \Pr(L = l \mid C) = \frac{1}{L_{\text{max}} - L_{\text{min}}}
  \]

- Node choice restrictions
  \[
  \Pr(M_x \mid L = l, C) = \frac{1}{P(N_{\text{mix}}, l)}
  \]

- Choose \( l \) out of the \( N_{\text{mix}} \) node available
- Choose a set \( I_{\text{set}}(M_x) \)

\[
\Pr(P_x \mid C) = \Pr(L = l \mid C) \cdot \Pr(M_x \mid L = l, C) \cdot I_{\text{set}}(M_x)
\]
Probabilistic model – Basic Constraints

- Unknown destinations

\[ L_{\text{min}} = 2 \quad L_{\text{max}} = 3 \]

\[
\Pr(P_x \mid C) = \left[ \sum_{l=L_{\text{obs}}}^{L_{\text{max}}} \Pr(L = l \mid C) \cdot \Pr(M_x \mid L = l, C) \cdot I_{\text{set}}(M_x) \right]
\]
Probabilistic model – More Constraints

- **Bridging**: users can only send through mixes they know

\[
\Pr( P_x | C ) = \Pr( L = l | C ) \cdot \Pr( M_x | L = l, C ) \cdot I_{set}( M_x ) \cdot I_{bridging}( M_x )
\]

- **Non-compliant clients** (with probability \( p_{cp} \))
  - Do not respect length restrictions \(( L_{\min, cp}, L_{\max, cp} )\)
  - Choose \( l \) out of the \( N_{\text{mix}} \) node available, allow repetitions

\[
\Pr( M_x | L = l, C, I_{cp}( Path )) = \frac{1}{P_r( N_{\text{mix}}, l )}
\]

\[
\Pr( Paths | C ) = \prod_x \Pr( P_x | C )
\]

\[
\Pr( Paths | C ) = \left[ \prod_{i \in P_{cp}} p_{cp} \Pr( P_i | C, I_{cp}( P_i )) \right] \cdot \left[ \prod_{j \in P_{cp}} (1 - p_{cp}) \Pr( P_j | C ) \right]
\]
Probabilistic model – More constraints

- **Social network information**
  - Assuming we know sending profiles $\Pr(\text{Sen}_x \rightarrow \text{Rec}_x)$
  $\Pr(P_x \mid C) = \Pr(L = l \mid C) \cdot \Pr(M_x \mid L = l, C) \cdot I_{set}(M_x) \cdot \Pr(\text{Sen}_x \rightarrow \text{Rec}_x)$

- **Other constraints**
  - Unknown origin
  - Dummies
  - Other mixing strategies
  - ....
Markov Chain Monte Carlo methods and traffic analysis
(The exciting part of the talk)
Markov Chain Monte Carlo

- Sample from a distribution difficult to sample from directly

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

- 3 Key advantages:
  - Requires generative model (we know how to compute it!)
  - Good estimation of errors
    - Not false positives and negatives
  - Systematic
Metropolis Hastings Algorithm

- Constructs a Markov Chain with stationary distribution \( \Pr(HS \mid O, C) \)
- Current state \( \xrightarrow{Q} \) Candidate state

1. Compute \( \alpha = \frac{\Pr(HS_{candidate})Q(HS_{candidate} \mid HS_{current})}{\Pr(HS_{current})Q(HS_{current} \mid HS_{candidate})} \)
2. If \( \alpha \geq 1 \)
   - \( HS_{current} = HS_{candidate} \) **Go!**
   - \( u \sim U(0,1) \)
   - If \( u \leq \alpha \)
     - Go with probability \( \alpha \)
   - Else \( HS_{current} = HS_{candidate} \)
   - \( HS_{current} = HS_{current} \)
Our sampler: $Q$ transition

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

$$\alpha = \frac{\Pr(\text{Paths}_{\text{candidate}})Q(\text{Paths}_{\text{candidate}} \mid \text{Paths}_{\text{current}})}{\Pr(\text{Paths}_{\text{current}})Q(\text{Paths}_{\text{current}} \mid \text{Paths}_{\text{candidate}})}$$

- **Transition $Q$: swap operation**

- **More complicated transitions for non-compliant clients**

C. Troncoso - UTA - Nov 2009
Iterations

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

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

- Consecutive samples dependant
- Sufficiently separated
Sampling error estimation

\[ P_1 \quad P_2 \quad P_3 \quad \ldots \]

\[ (A \to Q)? \quad 1 \quad 0 \quad 1 \quad \ldots \]

\[ \Pr( A \to Q) = \frac{\sum_i I_{A\to Q}(Paths_i)}{j} \]

Error estimation

\[ Pr[ \ I_{A\to Q}(P_1), I_{A\to Q}(P_2), I_{A\to Q}(P_3), \ldots \ | \ Pr( A \to Q) ] \]

\[ Pr[Pr( A \to Q) | I_{A\to Q}(P_1), I_{A\to Q}(P_2), I_{A\to Q}(P_3), \ldots] \]

Prior Beta(1,1) \sim \text{uniform}

\[ \Pr( A \to Q ) \sim \text{Beta} \left( \sum_{Paths} I_{A\to Q}(P_i) + 1, \sum_{Paths} I_{A\to Q}(P_i) + 1 \right) \]
Evaluation and results

(It actually works!)
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
   \[ \Pr(\text{Sen} \rightarrow \text{Rec}) \approx \frac{\sum_j I_{\text{Sen} \rightarrow \text{Rec}}(\text{Paths}_j)}{N} \]
5. Check if actually Sen chose Rec as receiver $I_{\text{Sen} \rightarrow \text{Rec}}(\text{network})$
6. Choose new network and go to 2
Example

- **Studying events with** $Pr(\text{Sen} \rightarrow \text{Rec}) = 0.4$

  - **Network 1**
    
    \[ I_{A \rightarrow B}(P_1) = 0; I_{A \rightarrow B}(P_2) = 1; I_{A \rightarrow B}(P_3) = 0; I_{A \rightarrow B}(P_4) = 0; I_{A \rightarrow B}(P_5) = 1 \]

    \[ \sum_j I_{A \rightarrow B}(P_j) \]

    \[ Pr(A \rightarrow B) \approx \frac{\sum_j I_{A \rightarrow B}(P_j)}{5} = 0.4; \quad I_{A \rightarrow B}(\text{Network 1}) = 0 \]

  - **Network 2**
    
    \[ Pr(X \rightarrow Y) = 0.4; \quad I_{X \rightarrow Y}(\text{Network 2}) = 0 \]

  - **Network 3**
    
    \[ Pr(X \rightarrow Y) = 0.4; \quad I_{X \rightarrow Y}(\text{Network 3}) = 1 \]

  - **Network 4**
    
    \[ Pr(X \rightarrow Y) = 0.4; \quad I_{X \rightarrow Y}(\text{Network 4}) = 1 \]

  - **Network 5**
    
    \[ Pr(X \rightarrow Y) = 0.4; \quad I_{X \rightarrow Y}(\text{Network 5}) = 0 \]

\[ Pr_{\text{sampled}}(X \rightarrow Y) = 0.4 \]

\[ Pr_{\text{empirical}}(X \rightarrow Y) \sim \text{Beta}(2 + 1, 3 + 1) \]
Results – compliant clients – 50 messages

Empirical probability

\[ E(I_{Sen \rightarrow Rec}(network)) \]

Predicted probability

\[ \sum_{Paths} I_{Sen \rightarrow Rec}(Paths_j) \]

C. Troncoso - UTA - Nov 2009
Results non compliant – 50 messages
Results – big networks

- It scales well as networks get larger
- As expected mix networks offer good protection
## Performance – RAM usage

<table>
<thead>
<tr>
<th>Nmix</th>
<th>t</th>
<th>Nmsg</th>
<th>Samples</th>
<th>RAM(Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>500</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>50</td>
<td>500</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>50</td>
<td>500</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1000</td>
<td>500</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>10000</td>
<td>500</td>
<td>125</td>
</tr>
</tbody>
</table>

- Size of network and population
- Results are kept in memory during simulation
## Performance – Running time

<table>
<thead>
<tr>
<th>Nmix</th>
<th>t</th>
<th>Nmsg</th>
<th>iter</th>
<th>Full analysis (min)</th>
<th>One sample (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>6011</td>
<td>4.24</td>
<td>509.12</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>50</td>
<td>6011</td>
<td>4.80</td>
<td>576.42</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>100</td>
<td>7011</td>
<td>5.34</td>
<td>641.28</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1000</td>
<td>7011</td>
<td>5.97</td>
<td>706.12</td>
</tr>
</tbody>
</table>

- Operations should be $O(1)$
  - Writing of the results on a file
  - Different number of iterations
Applications

- Evaluation information theoretic metrics for anonymity
  \[ H = - \sum_{R_i} P(A \rightarrow R_i | O, C) \cdot \log P(A \rightarrow R_i | O, C) \]

- 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: incorporates most attacks
  - Non-compliant clients

- Monte Carlo Markov Chain methods to extract marginal probabilities
  - Systematic
  - Only generative model needed

- Future work:
  - Model more constraints
  - Added value?
Time for questions

More info

- **Vida: How to use Bayesian inference to de-anonymize persistent communications.** George Danezis and Carmela Troncoso. Privacy Enhancing Technologies Symposium 2009

- **The Bayesian analysis of mix networks.** Carmela Troncoso and George Danezis. 16th ACM Conference on Computer and Communications Security 2009

- **The Application of Bayesian Inference to Traffic analysis.** Carmela Troncoso and George Danezis Microsoft Technical Report

Carmela.Troncoso@esat.kuleuven.be

http://homes.esat.kuleuven.be/~ctroncos/

C. Troncoso - UTA - Nov 2009
Bayes theorem

\[ \Pr(O, HS \mid C) = \Pr(HS \mid O, C) \cdot \Pr(O \mid C) \]

\[ \Pr(O, HS \mid C) = \Pr(O \mid HS, C) \cdot \Pr(HS \mid C) \]

\[ \Pr(HS \mid O, C) = \frac{\Pr(O \mid HS, C) \cdot \Pr(HS \mid C)}{\Pr(O \mid C)} = \frac{\Pr(O \mid HS, C) \cdot \Pr(HS \mid C)}{\sum_{HS} \Pr(HS, O \mid C)} \]

Joint probability:

\[ \Pr(X, Y) = \Pr(X \mid Y) \cdot \Pr(Y) = \Pr(Y \mid X) \cdot \Pr(X) \]
Error estimation: the Beta function

- We need to specify the **prior knowledge** \( \Pr(\theta) \) \( \Pr(HS \mid C) \)
  - expresses our uncertainty
  - conforms to the nature of the parameter, i.e. is continuous but bounded between 0 and 1

- A convenient choice is the Beta distribution

\[
P(\theta) = \text{Beta}(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1}
\]

![Graphs of Beta distributions with different parameters](image)
Error estimation: the Beta function

- Combining a beta prior with the binomial likelihood gives a posterior distribution

\[
p(\theta \mid \text{total, successes}) = p(\text{successes} \mid \theta, \text{total}) \cdot p(\theta) = \theta^{\text{successes} + a - 1} (1 - \theta)^{\text{total} - \text{successes} + b - 1} \\
\propto \text{Beta}(\text{successes} + a, \text{total} - \text{successes} + b)
\]

Prior knowledge