Inferring Services over Encrypted Web Flows

電子情報通信学会　総合大会 (EB-7-71)
2014年3月20日

森　達哉 (早稲田大学), 井上　武 (JST ERATO)

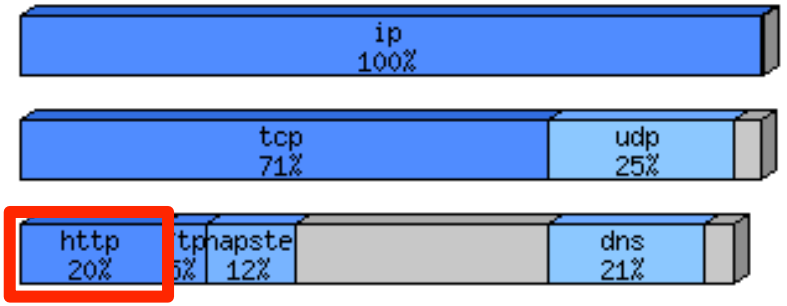
NSL
Networked Systems Laboratory

背景(1) web 全盛の時代

インターネットトラフィックの変化

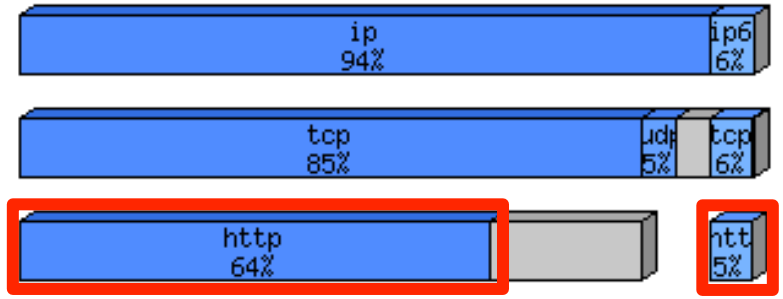
※WIDE Mawi Project http://mawi.wide.ad.jp, samplepoint B, F のグラフを引用

2002/12/1



P2Pが全盛

2012/12/1

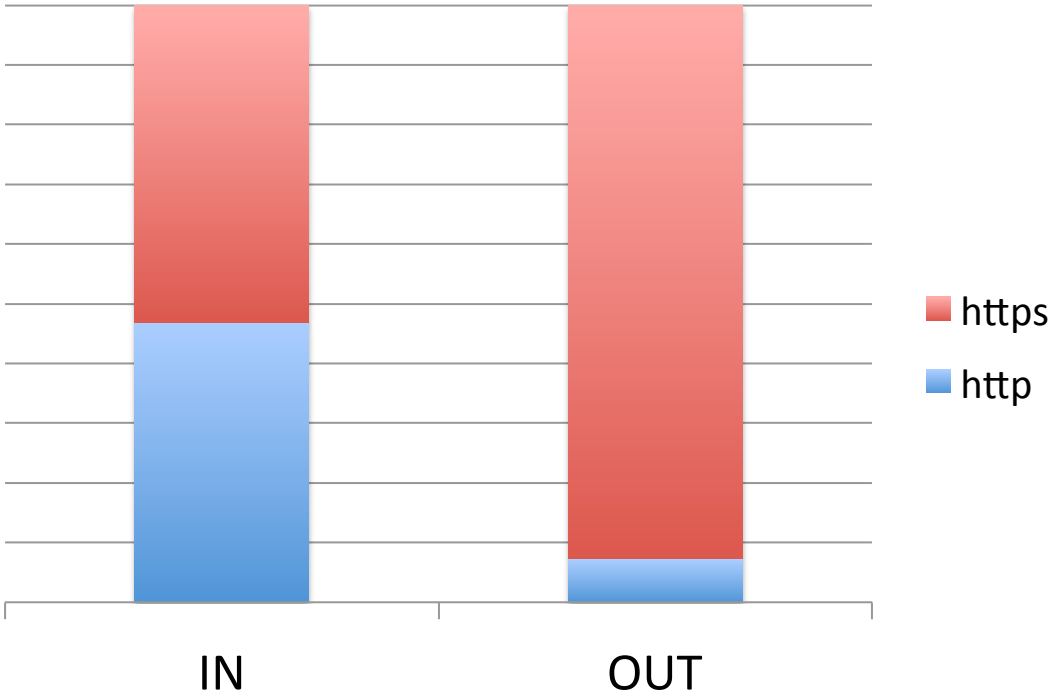


Webが全盛

インターネット上の多くのサービスが Web(http) に移行

背景(2) web 通信の暗号化

被験者延べ14人分のAndroid の通信を6時間計測した結果を集計



SSL/TLS によって暗号化されているHTTP通信(HTTPS)の割合が高い
⇒ソーシャルネットワークやオンラインストレージ等の
パーソナルな情報を送受信するアプリケーションの増加

NSL
Networked Systems Laboratory

インターネットトラフィック計測の目的と課題

- 目的: トラフィック内訳の把握
  - 制御対象把握: 帯域の枯渇時に何をフィルター対象とすればよいか
    - 単に HTTP では範囲が広すぎる
  - 需要の把握: どのようなアプリケーションに対する要求が高いのか？

  →アーキテクチャの検討（キャッシュの設置やWAN 最適化等）

  ※いずれもざっくりとした統計があれば十分

- 課題: 既存アプローチの限界を克服
  - ポート番号では情報量が少なすぎる
  - サーバのIPアドレスでは不十分（次スライド）
  - DPI の限界: SSL/TLS 利用時に HTTP ヘッダが参照できない

NSL
Networked Systems Laboratory

サーバのIPアドレスでは不十分な例

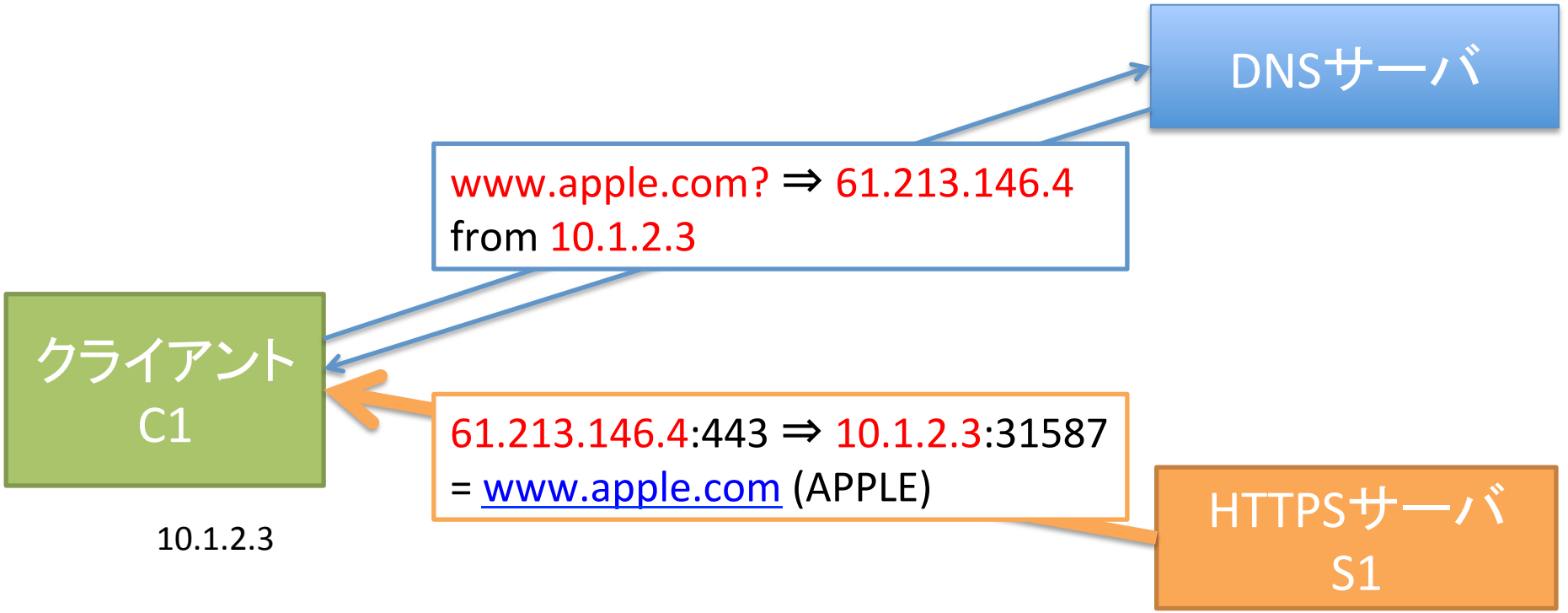
- 逆引きを設定していないケースが存在
- 複数の FQDN がひとつのIPアドレスで使われているケース (ホスティングやCDN 等)

本研究のゴール

- インターネットトラフィックを生成しているWebアプリケーション(＝サービス)を把握する
  - 特に暗号化 Web 通信を対象とする
  - 100% の精度を目指すものではない
    - 原理的に困難　⇒　ざっくりとした統計の把握

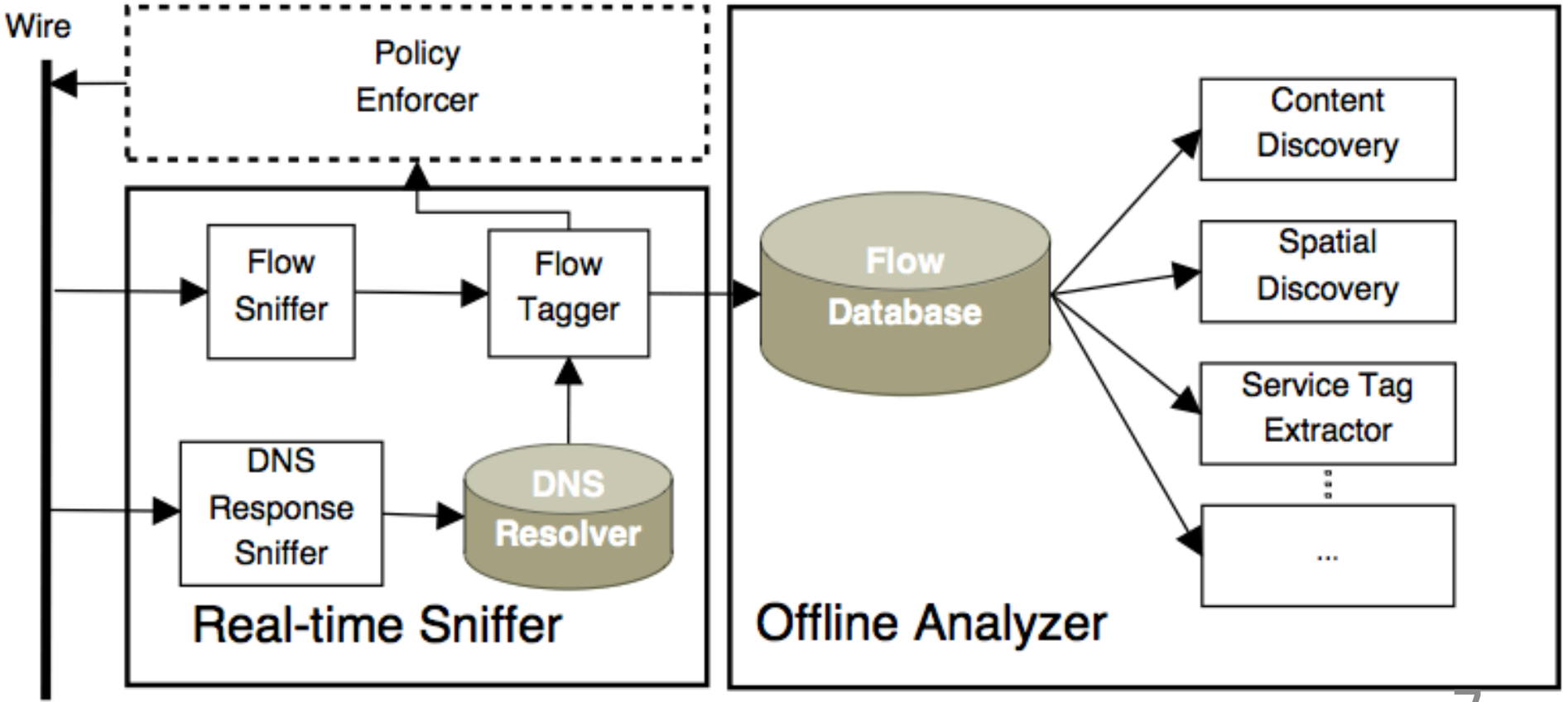
DNS クエリを用いたアプローチ

通信の開始に先立って クライアントからDNS の名前解決が発生するので，その情報を用いればホスト名(FQDN)はわかる



既存研究: DN Hunter

- Bermudez et al., “DNS to the Rescue: Discerning Content and Services in a Tangled Web”, ACM IMC 2012



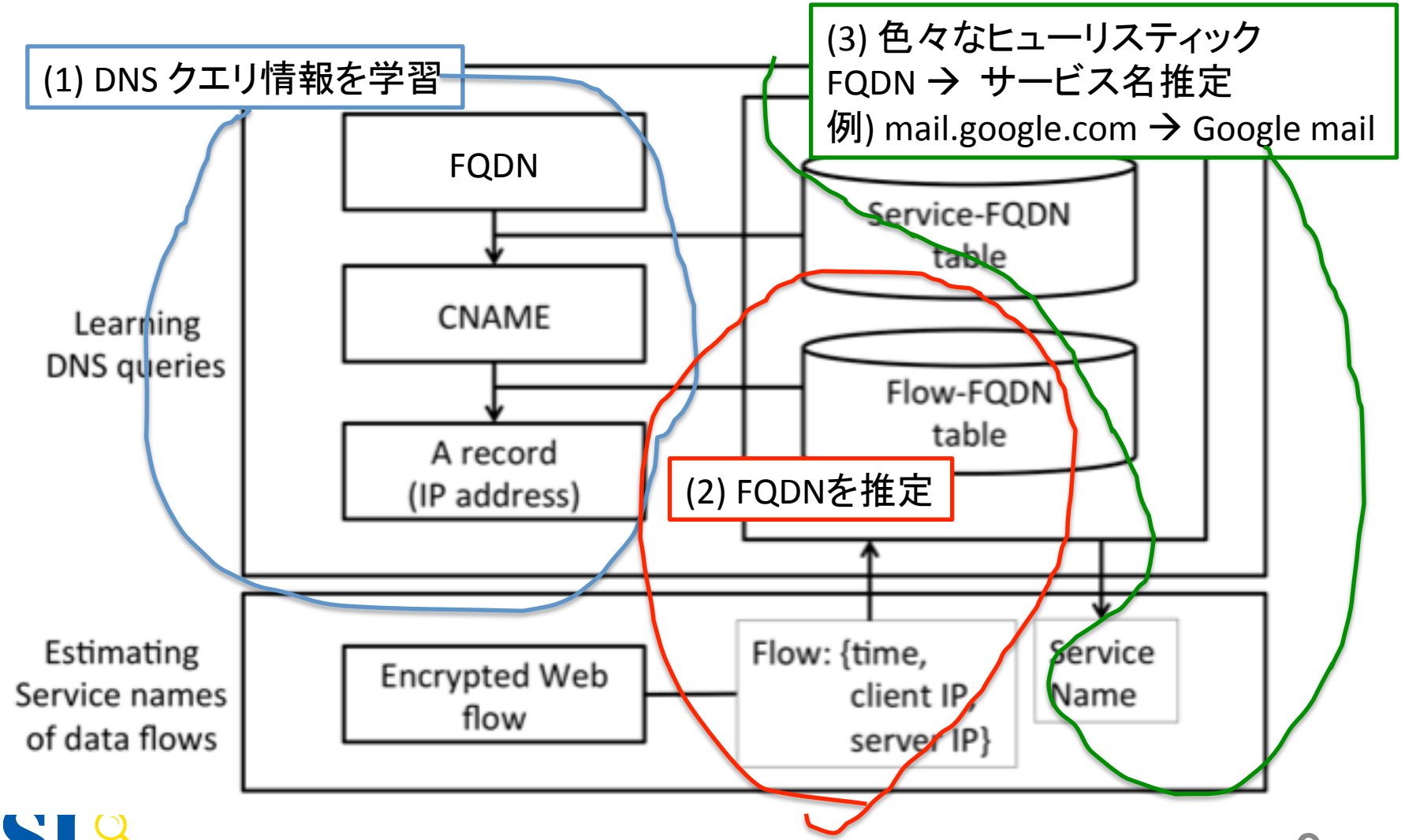
※図面は上記論文より引用

既存研究と提案手法の比較

	分散計測対応	統計的推定
DN Hunter	△	× (見たものがすべて)
提案手法	○	○ (見てないものも推定)

ヒットしなかった場合においても過去の観測をもとに
統計的に推定する点が新規なアイディア

提案手法の全体像



NSL
Networked Systems Laboratory

提案方式の概要 (1): 学習

- サーバIPアドレス: s
- クライアントIPアドレス: c
- 時刻: t (秒単位とする)
- DNS A レコード query における FQDN: N

としたとき

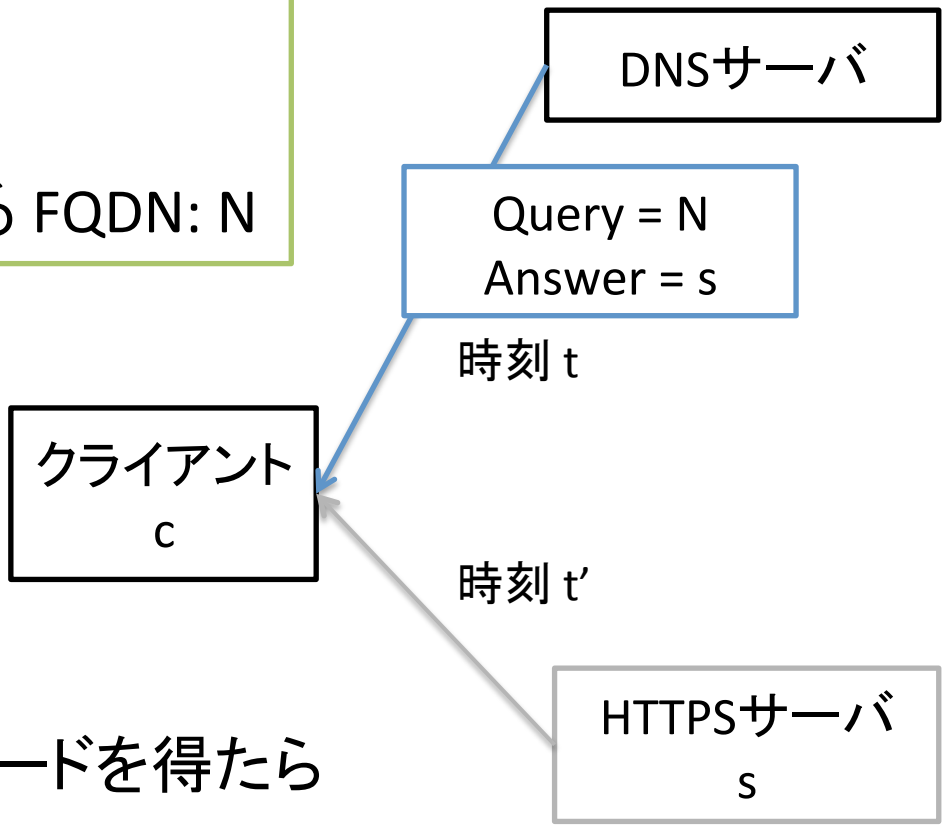
{s,c,t} → N

{s,c} → N

{s} → N

のすべてを連想配列に登録

※Query response で複数 A レコードを得たら
すべて登録



提案方式の概要 (2): 推定

- 推定の対象となる HTTPS フローの {s, c, t'} を抽出
- イグザクトマッチ
  - {s, c, t'} に対する N を出力
- 時間ずらし検索
  - Exact match が失敗した場合，{s, c, t'} の t' を減らしながら検索 (t' = t', t'-1, t'-2, …, t'-m)
    - DNS クエリが HTTP 通信開始に先立って発生するケースがある
- 統計的推定 (MAP)
  - 上記いずれも失敗した場合，{s,c} あるいは {s} のみを使って尤もらしいFQDN を統計的に推定（最大事後確率推定）

最大事後確率推定 (MAP)

●{s, c} のクエリに対する応答 N={n1, n2, …,} の内，尤もらしいFQDNを下記のように推定

$$\hat{n} = \arg\max_{n \in \mathbf{N}} P(n|s,c) = \arg\max_{n \in \mathbf{N}} P(s,c|n)P(n)$$

ある FQDN に対して (s, c) の組み合わせが出現する確率

ある FQDN の出現頻度
サービスの人気度

●{s, c} のクエリに対する応答がない場合はクエリを {s} として同様に推定

$$\hat{n} = \arg\max_{n \in \mathbf{N}} P(n|s) = \arg\max_{n \in \mathbf{N}} P(s|n)P(n)$$

※ 最大事後確率の計算にはいくつかのバリエーションが考えられる

提案方式の概要 (3): サービス推定

- FQDN を形成する文字列の特徴にもとづき，サービスを推定

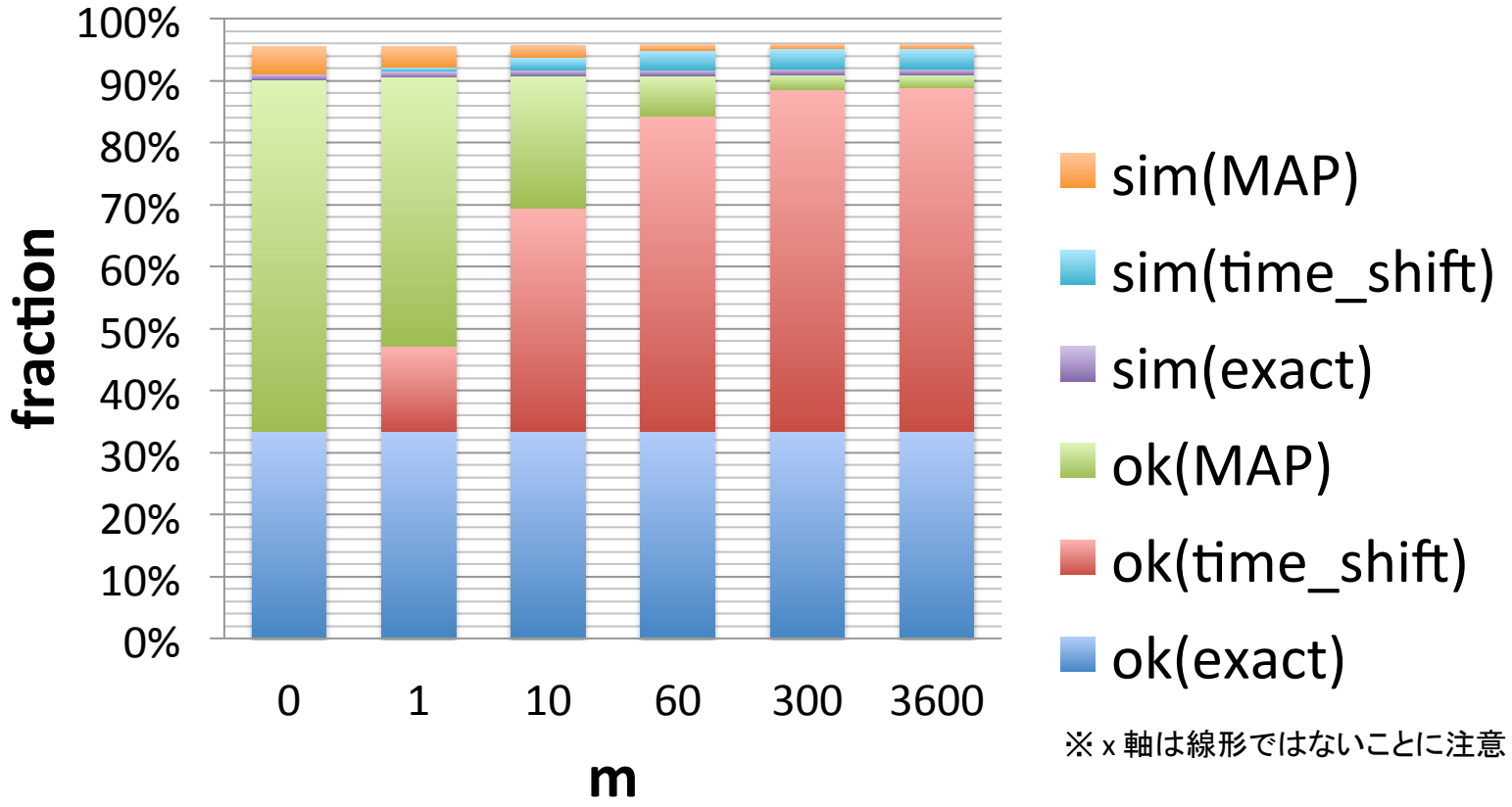
- Public suffix を抽出
  - www.ieice.org の public suffix = ieice.org
- 残りの文字列に対して特徴的な文字列の有無を判定
  - mail, blog, platform, ad, 等

性能評価実験

- 約2,000の端末が通信をしている回線を計測
- 暗号化されていない HTTP 通信を利用
  - Request header から真の FQDN を抽出可能
  - HTTP リクエスト数: 30084
  - 時間 = 約4000秒
- DNSクエリ
  - 上記 HTTP 通信の時間帯を含む46000秒
  - 約10万クエリ

- 今回の実験評価では推定結果を以下のように場合分けをする
  - 推定FQDNと真のFQDN が完全一致 (OK)
  - Public suffix が一致 (SIM)
  - それ以外 (NG)

パラメタ m と検索精度

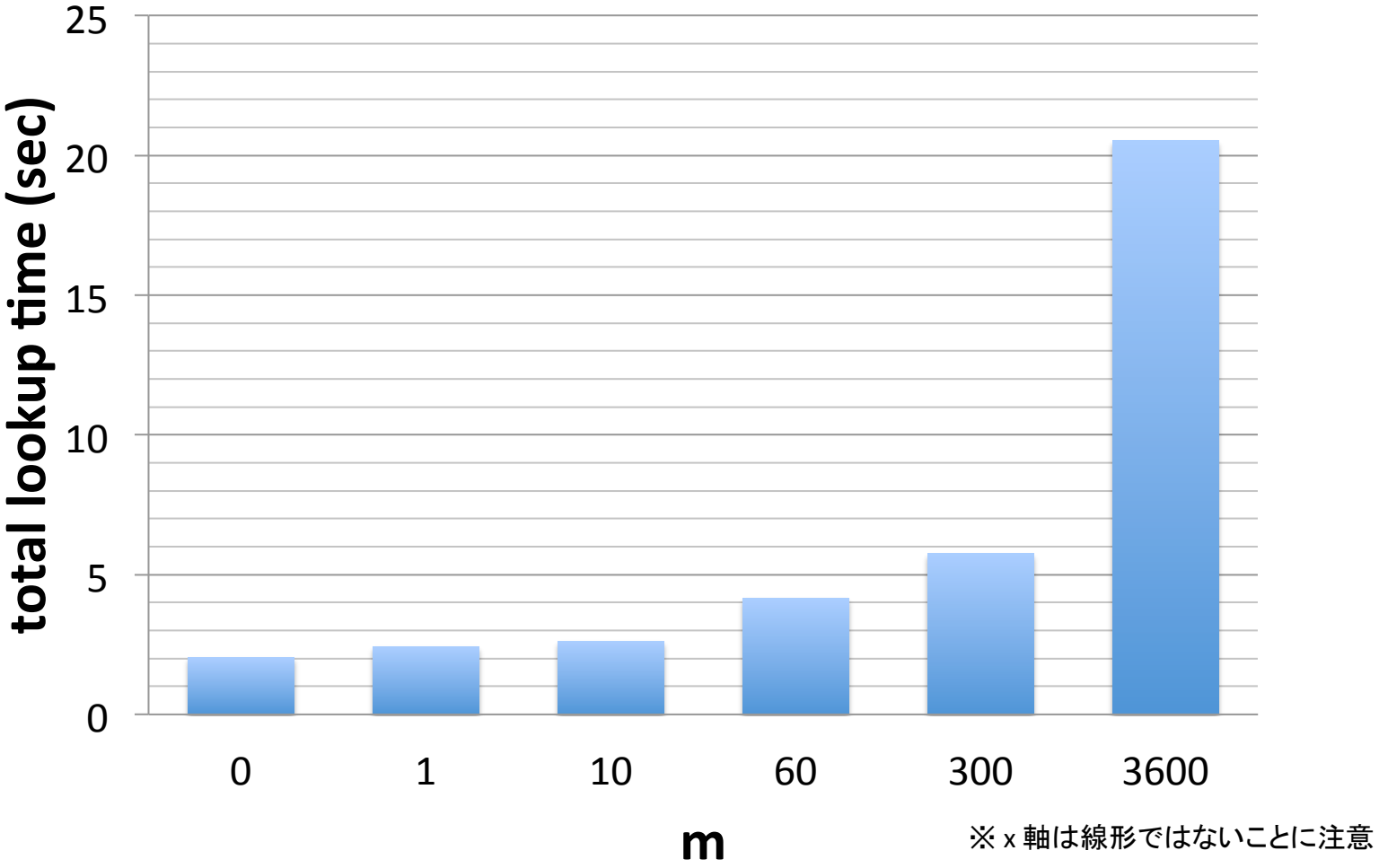


※ x 軸は線形ではないことに注意

m に寄らずOK は 90%, SIM が 5% 程度の精度
m = 0 (時間ずらし検索なし)でも精度は良い (MAPの推定結果が主)
m を増やすと時間ずらし検索の推定結果が主となるが結果は同様

NSL
Networked Systems Laboratory

パラメタ m と検索コスト



m の増加とともに検索コストが高まる

エラーの原因

- 観測データ不足
  - 観測期間内に観測したweb サーバの IP アドレスを含むDNSクエリが発生しない場合は推定のしようがない

- 同時刻に発生する {s,c,t} のタプル
  - 現在は t を秒単位にまるめているが，もう少し細かい時間分解能が必要

例) googleads.g.doubleclick.net と
pagead2.googlesyndication.com が同一の {s,c,t} を持つ

  - どちらも CNAME = pagead46.l.doubleclick.net

まとめと今後の課題

- 過去の DNS クエリを参照して暗号化 Web 通信のサービスを推定する手法を提案
- 精度は完全一致が 90%，public suffix 一致が5%程度
- MAP を使うことで検索コストを短縮可能

- 課題 (1) 精度向上
  - 長時間データの取り込み，時間分解能，ヒューリスティックの開発
- 課題 (2) スケーラビリティの確立
  - 膨大な通信ログへの対応
  - 過去データの蓄積

謝辞

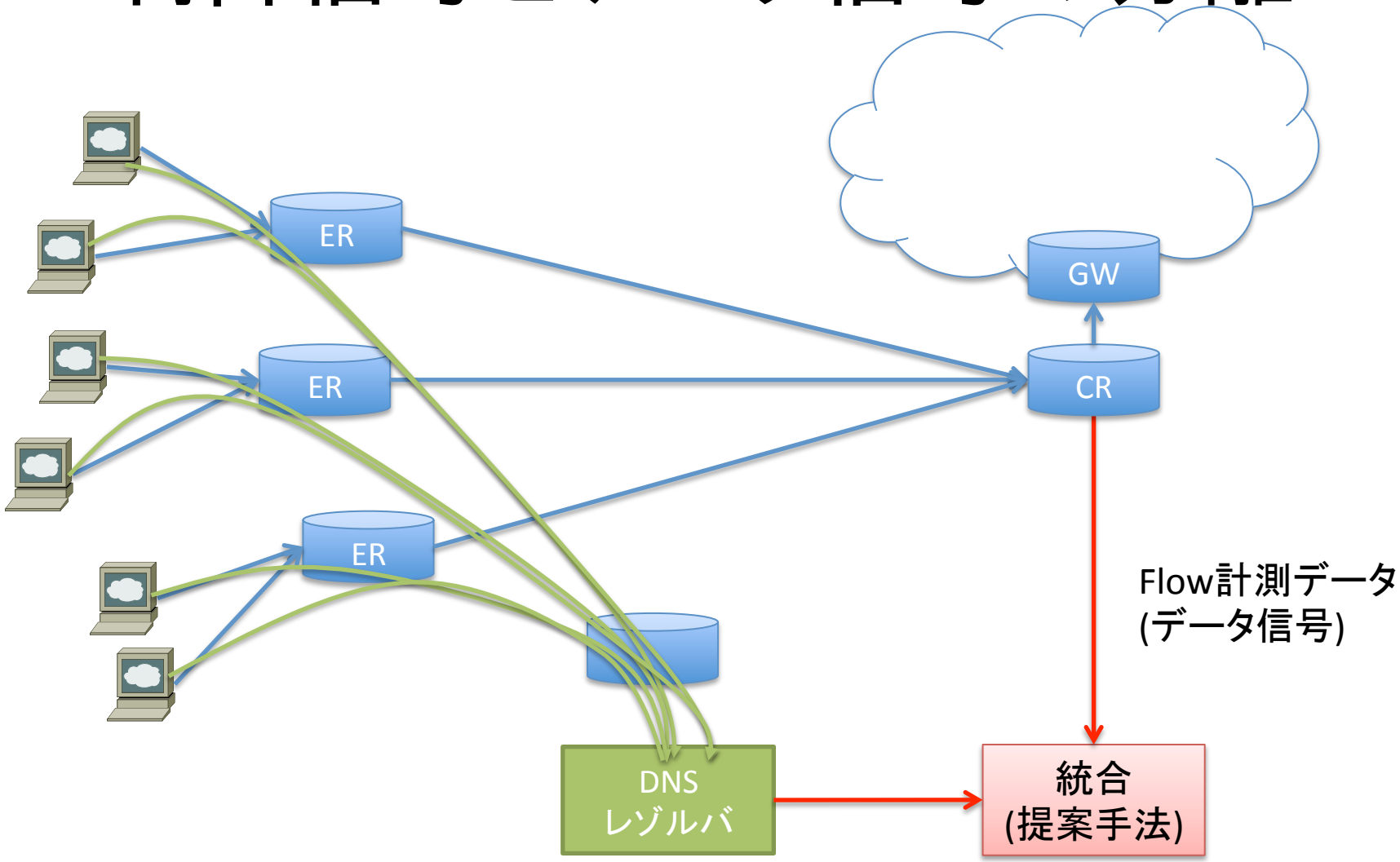
- 本研究の一部はJSPS科研費（25880020，代表　森達哉）の助成を受けたものです．

- 本研究に関して議論いただいたNTTネットワーク基盤技術研究所の石橋主幹研究員，佐藤研究員，下田研究員に感謝します．

SSL/TLS の利用

- SSL/TLS で暗号化された Web 通信に関しては公開鍵証明書に記載された CommonName (URL の FQDN と一致) の利用が可能であるが，FQDN の把握には不十分
  - DN-Hunter 論文 (ACM IMC 2012) での分析結果
    - CN = FQDN : 18%
    - ワイルドカード証明書: 19%
    - まったく異なる証明書(?): 40%
    - 証明書無し: 23%

制御信号とデータ信号の分離



原理的にクエリを観測できないケース

- ブラウザの DNS キャッシュ機能やユーザエンドのルータに実装された DNS キャッシュサーバにより，観測地点で DNS クエリが観測できない
- 稀にIP アドレス直打ちのケースがある
- 端末のIPアドレス再割当てや退去