Anonymous Data Collection System with Mediators

Hiromi Arai[1], Keita Emura[2], Takahiro Matsuda[3]

1. The University of Tokyo, Japan
2. National Institute of Information and Communications Technology (NICT), Japan
3. National Institute of Advanced Industrial Science and Technology (AIST), Japan

Contents

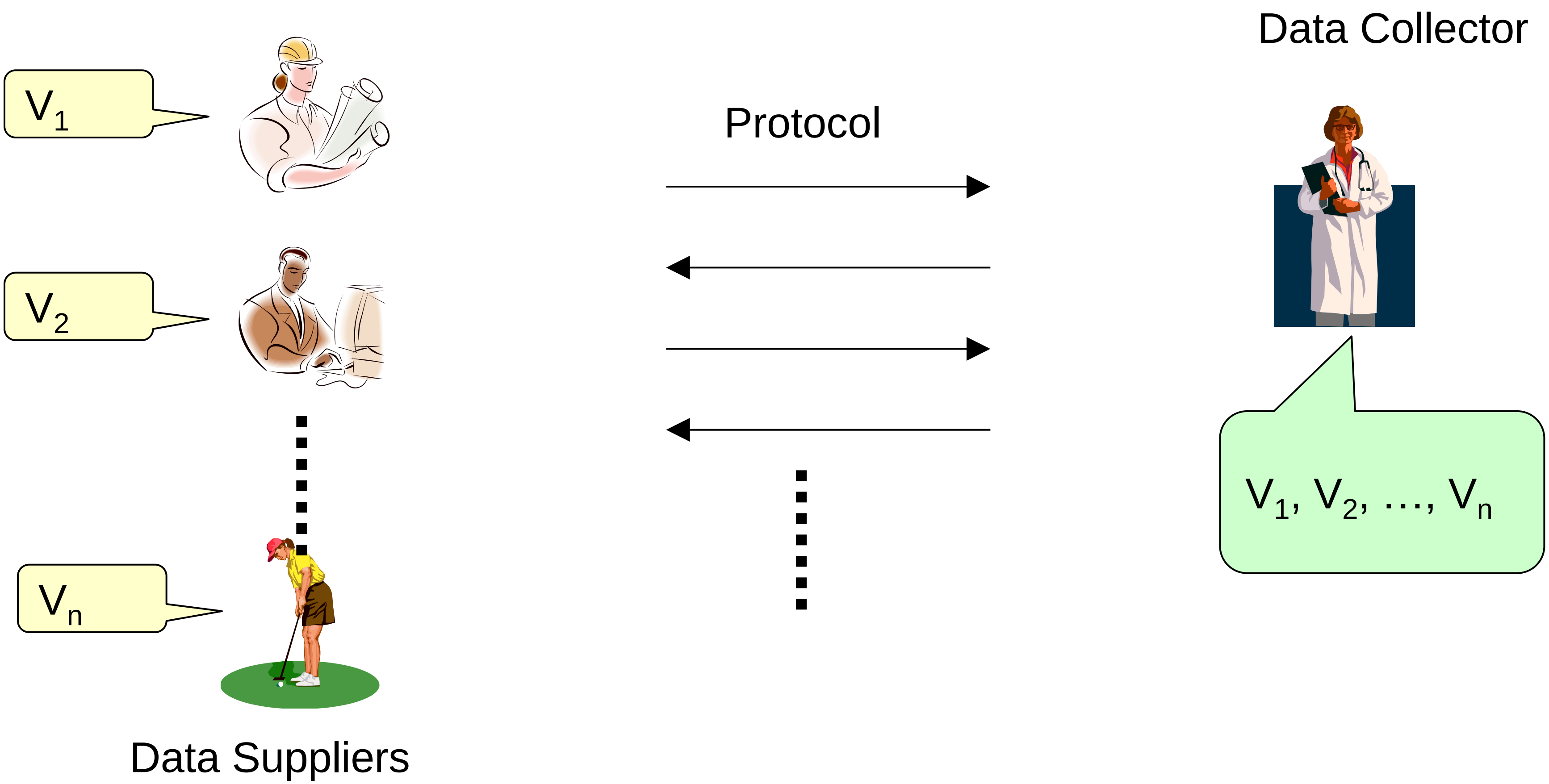
- Research Background
  - Secure Data Collection
- Our Contribution
  - Secure Data Collection with Mediators
    - Delegate a data collection task to Mediators in a “secure” way
- Proposed System
  - A generic construction of the system from restrictive public key encryption (RPKE)
- Efficiency Estimation
- Conclusion

Privacy-preserving Data Mining

- Sensitive data is treated for a constellation of purposes:
  - e.g., establishing the presence or absence of causal association among certain diseases.
- Statistics of sensitive data need to be computed
  - secure computation, differential privacy, k-anonymity, etc.

- How to collect sensitive data with concerning privacy in the first place?

Data Collection



Secure Data Collection

- Zhiqiang Yang, Sheng Zhong and Rebecca N. Wright: Anonymity-preserving data collection, KDD 2005.
- Justin Brickell and Vitaly Shmatikov: Efficient anonymity-preserving data collection, KDD 2006.
  - By employing public key encryption (PKE) and digital signature as its building blocks.
  - Two entities: a data collector and data suppliers
- Mafruz Zaman Ashrafi and See-Kiong Ng: Collusion-resistant anonymous data collection method, KDD 2009

The Brickell-Shmatikov System

- PKE, digital signature
- Two entities: a data collector and data suppliers
- Security
  - Anonymity w.r.t. collusion resistance
    - Anonymity holds even if all data suppliers, except two honest ones, collude with each other
  - Integrity
    - If the protocol does not abort, then all honest suppliers' data are contained in the collection result.
  - Confidentiality
    - If a data collector is honest, then no honest suppliers' data is revealed to any dishonest data supplier.

The Brickell-Shmatikov System

1. No formal cryptographic definitions were given（about Integrity and Confidentiality）in their work, though cryptographic tools are employed in their systems.

2. All data suppliers are required to be on-line during the data collection procedure（anonymization and verification）, and the number interaction between the data collector and data suppliers is linear in the number of data suppliers.

3. Large Ciphertext Overhead: One data is sequentially encrypted n-times (n=#Suppliers, $|C|=O(n^2)$)
In total, comm. Overhead: $O(n^3)$.

- If a data collector is honest, then no honest suppliers' data is revealed to any dishonest data supplier.

How to reduce the cost?

- One approach is using <u>**mediators.**</u>
  - The data collector can delegate the data collection task to them.
  - In many practical situations in which sensitive data is collected, <u>**the data collector does not necessarily have to identify**</u> data suppliers.
    - Managing identity table courses a risk for its <u>**exposure**</u> and unnecessary data should not be managed as much as possible.
  - Of course, the data collector should not reveal data itself to mediators.

Naïve Approach

- Using PKE
  - the data collector has a public key and data suppliers encrypt their data using the public key
  - the data collector checks collected data after decrypting these ciphertexts.
- This does not make sense since mediators do nothing and the cost for data collection is not reduced.

- For reducing the costs of the data collector, giving format-check capabilities (e.g., check whether data belongs to a certain range) to mediators is effective.

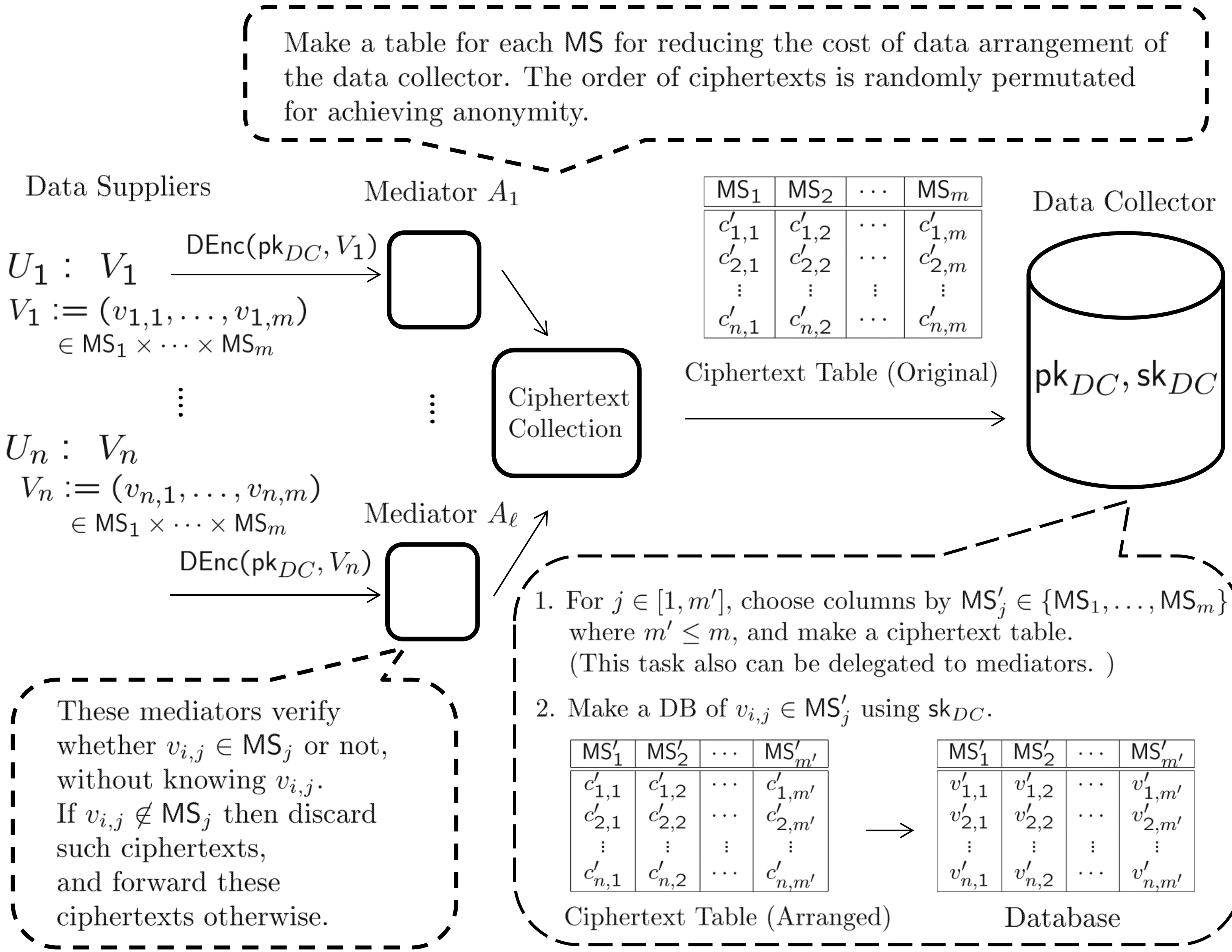
Alternative Solution and its Limitation

- Should symmetric key cryptography be employed for the fast decryption?
  - The data collector is required to run a key exchange protocol for **each** data suppliers.
    - Requiring interaction or similar cost of PKE
  - Hybrid Encryption?
    - The decapsulation cost is almost the same as that of the decryption cost of usual PKE.

Our Contribution

- Anonymous data collection system **with mediators**
  - The data collector can delegate data collection and data arrangement tasks to mediators **in a secure way** so that no mediator can know (unallowable information of) actual data.
  - Mediators can **check a data format without knowing data itself** so that data belongs to a certain range
    - age, gender, disease and so on
    - can sort out (encrypted) data by regarding a range as a quasi-identifier.
  - There is **no interaction between data suppliers and data collector**, i.e. no data supplier is required to be on-line during the data collection procedure.
  - **Ciphertext Overhead: O(1)** (in total, comm. Overhead O(n))
  - Give **formal cryptographic security definitions** (semantic security, anonymity, and format-check soundness
    - Provably secure

Brief Description



Brief Description

Example

MS_1:=10s,…, MS_9:=90s and MS_{10}:=Disease

Look like a k-anonymized table

Mediator

Age				Disease
MS_1	MS_2	$\cdots$	MS_9	MS_{10}
—	$c'_{1,2}$	$\cdots$	-	$c'_{1,10}$
—	—	$\cdots$	$c'_{2,9}$	$c'_{2,10}$
⋮	⋮	⋮	⋮	⋮
$c'_{n,1}$	—	$\cdots$	—	$c'_{n,10}$

Decryption

Data Collector

Age				Disease
MS_1	MS_2	$\cdots$	MS_9	MS_{10}
—	23	$\cdots$	-	Flu
—	—	$\cdots$	94	Diabetes
⋮	⋮	⋮	⋮	⋮
18	—	$\cdots$	—	NoDisease

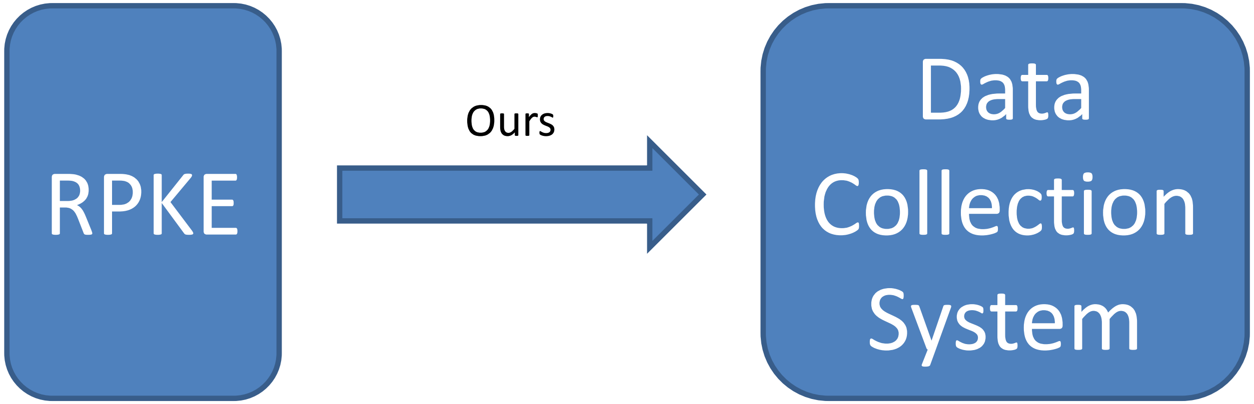
and forward these
ciphertexts otherwise.

Ciphertext Table (Arranged)

Database

Our Construction

- Propose a <u>**generic**</u> construction from restrictive public key encryption (RPKE)
  - RPKE: PKE with non-interactive range proof and decryption



System syntax

A secure data collection system $\mathcal{SDCS}$ consists of five algorithms (KeyGen, DEnc, FormatCheck, TableGen, DDec):

$\mathsf{KeyGen}(1^\kappa)$

$\mathcal{MS} := (\mathsf{MS}_1, \ldots, \mathsf{MS}_m)$: a set of message spaces
pk_{DC}: a public key
sk_{DC}: a secret key

This algorithm is supposed to be run by Data collector.

$\mathsf{DEnc}(\mathsf{pk}_{DC}, V = (v_1, \ldots, v_m) \in \mathsf{MS}_1 \times \cdots \times \mathsf{MS}_m, \ )$

$C_D := (c_1, \ldots, c_m)$: a ciphertext

This algorithm is supposed to be run by each Data supplier.

$\mathsf{FormatCheck}(\mathsf{pk}_{DC}, C_D, \ )$

f-index $:= \{1, \ldots, m\}$

for each $j \in [1, m]$ change j-th element of f-index to ϵ if the corresponding data $v_j \notin \mathsf{MS}_j$

$\mathsf{TableGen}(\mathcal{MS}, \mathsf{pk}_{DC}, (C_{D,i})_{i=1}^{n})$

$C_{D,\phi(i)}$ where $\phi : [1, n] \to [1, n]$ is a random permutation

These algorithms are supposed to be run by Mediator.

$\mathsf{DDec}(\mathsf{pk}_{DC}, \mathsf{sk}_{DC}, C_D)$

$(v_1, \ldots, v_m)$ or $\perp$

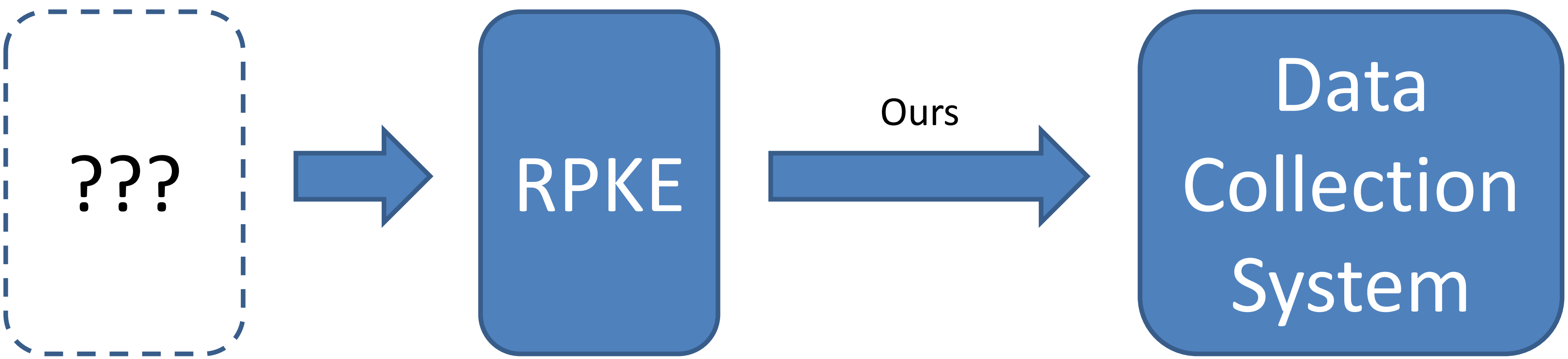
This algorithm is supposed to be run by Data collector.

Security Requirements

- Anonymity
  - Guarantee that Data collector obtains no information of data provided by Data suppliers.
    - Suppose that Data suppliers to collude with others, except two honest Data suppliers
      - (collusion resistance as in Brickell and Shmatikov)
- Semantic security
  - Guarantee that no information of data v is revealed from a ciphertext.
    - No Mediator can know v, except the fact that v belongs to some message space MS.
- Format-check soundness
  - Guarantee that for all C_j in table, if C_j passes the check by FormatCheck, the decryption result of C_j belongs to MS_j.

Our Construction

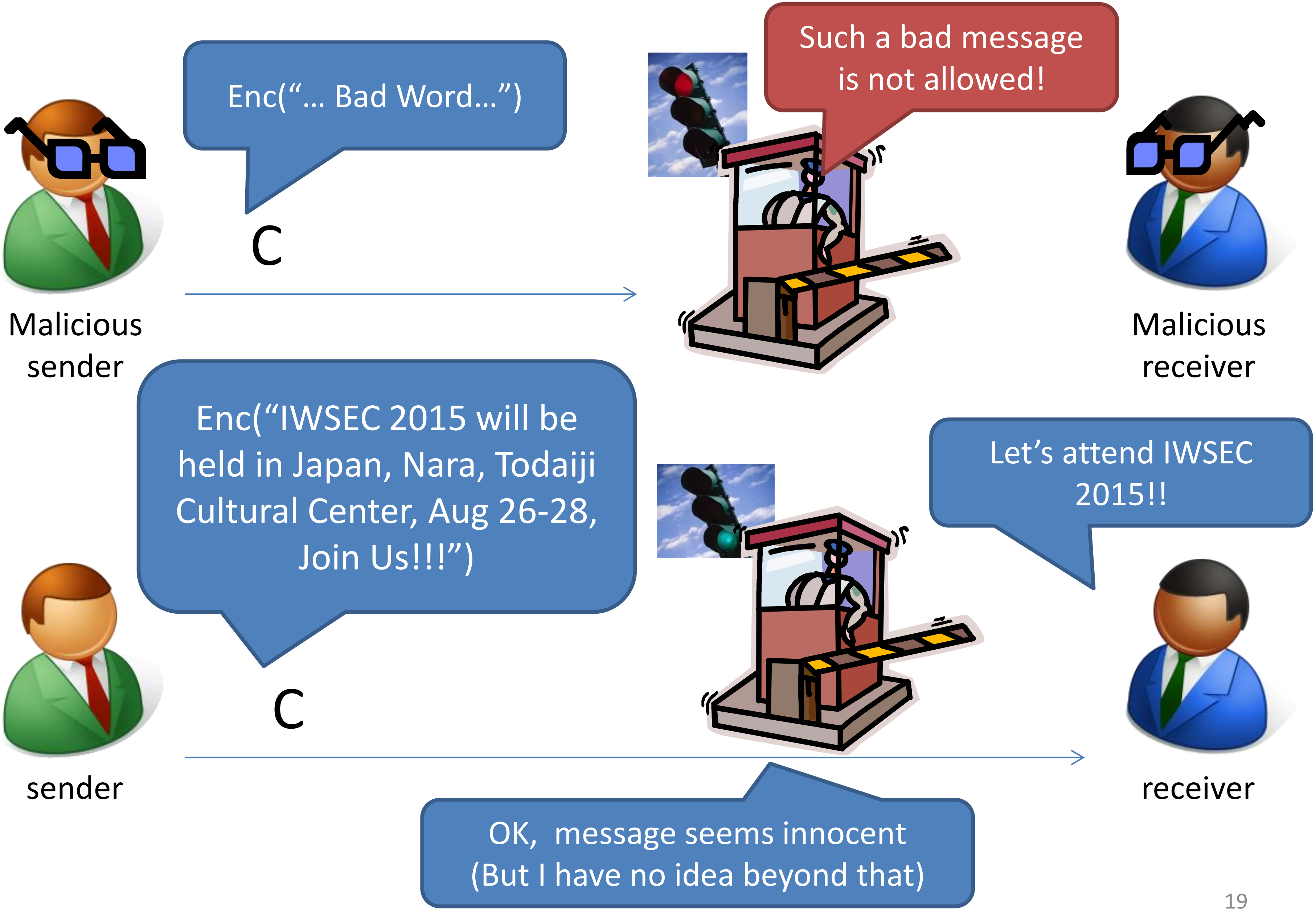
- Propose a **generic** construction from restrictive public key encryption (RPKE)
  - RPKE: PKE with non-interactive range proof and decryption



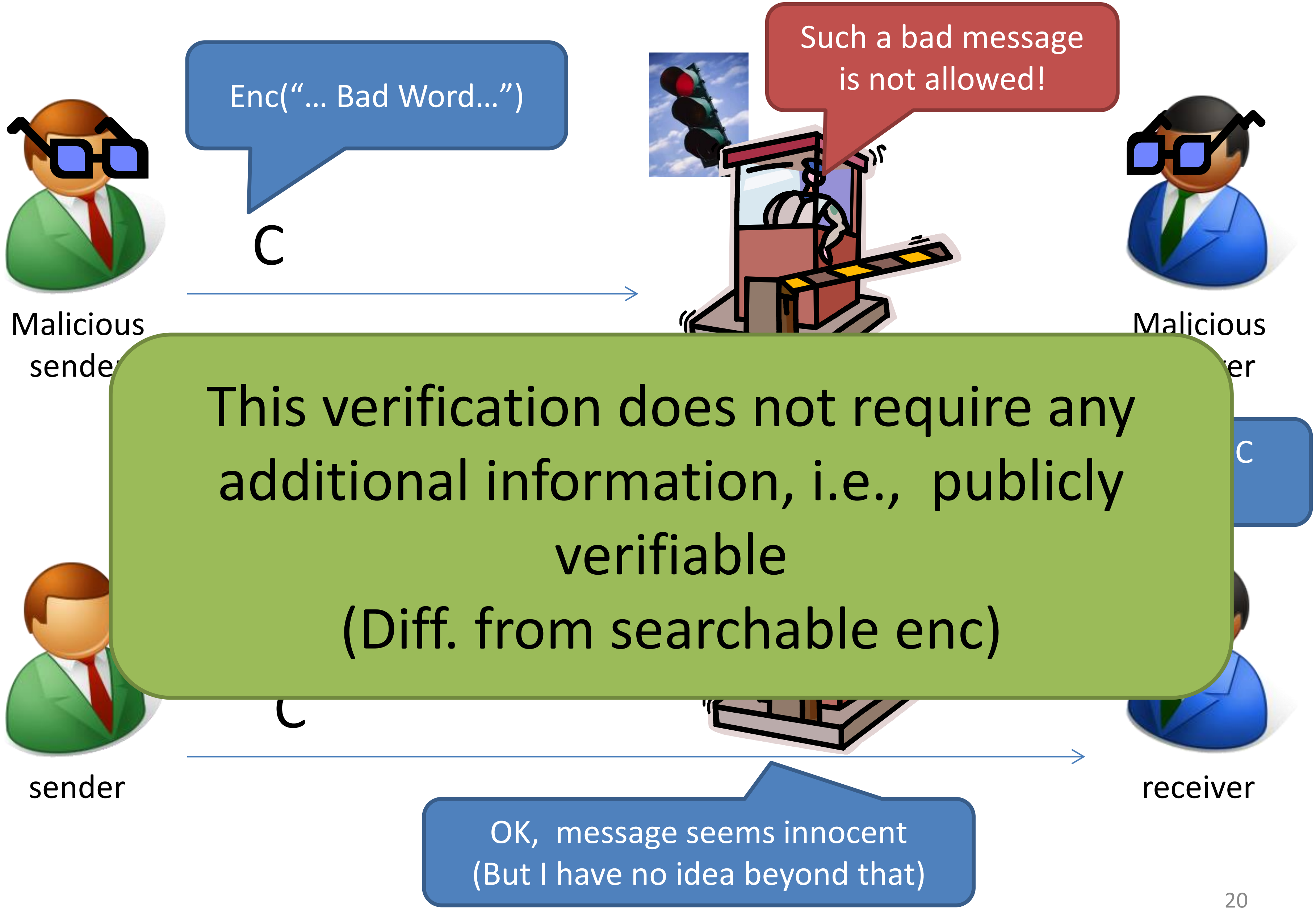
The Sakai et al. RPKE scheme

- Yusuke Sakai, Keita Emura, Goichiro Hanaoka, Yutaka Kawai and Kazumasa Omote: Towards Restricting Plaintext Space in Public Key Encryption, IWSEC 2011
  - Full version: IEICE trans. 2013
- Suitably restrict a plaintext space of PKE
  - Apply the revocation technique of group signature

Restrictive PKE



Restrictive PKE



Drawback of the Sakai et al. construction

- Support range proof and decryption simultaneously
  - Solving the DL problem for decryption
    - Lifted ElGamal-type construction
    - Message spaces are required to be sufficiently small

- We propose a **generic** construction of the data collection system from any RPKE
  - Constructing an efficient RPKE scheme is an interesting future work of this paper

Efficiency Estimation

- The Sakai et al. RPKE scheme
- The PBC library
  - We compiled the benchmark program with gcc 4.4.7 and run it on a 3.10-GHz Intel(R) Xeon(R) Processor E3-1220 64-bits PC (CentOS release 6.4) with 8 GB memory.
  - We use a (Type A) curve $y^2=x^3+x$.
    - A base group element is 512 bits, and a target group element is 1024 bits.

Efficiency Estimation

Running Time (Basic Operations)

Operation	Time(msec)
Pairing	1.146
Exp. ($\mathbb{G}$)	1.727
Exp. ($\mathbb{G}_T$)	0.149
Exp. ($\mathbb{G}'$)	0.617

Running Time (Algorithms)

Algorithm	Time(msec)	Entity
DEnc	$59.822m$	Data supplier
FormatCheck	$68.708m/\ell$	Mediator
DDec	$0.617m'$	Data collector

M: #Message spapces

ℓ: #Mediators

m': #Message spapces involved in the current data mining

Further Extension

- More Flexible Systems
  - In our system syntax, message spaces are fixed in the setup phase
  - A message-space setup algorithm is defined in the syntax of RPKE.
    - Message spaces can be changed (without full re-setup) for each data mining/data processing
      - by executing the message-space setup algorithm again, and Data Suppliers use the new public key.

Future Work

- Privacy-Preserved Outcome
  - Mediators can obtain k-type anonymized table only
    - But this k-type anonymization might be improved by considering l-diversity [MGKV06], t-closeness [LLV07] and p-sensitivity [TCM07] etc.
- Malicious adversarial model
  - Mediators and Data suppliers are modeled as semi-honest parities
- Efficient RPKE scheme
  - Without solving DL problem
- Relation from other techniques:
  - k-concealment (k-anonymity with comp. indistinguishability) [TMG12]
  - re-identification of k-anonymized data sets [StokesT12]
    - Our system supports decryption

Conclusion

- Anonymous data collection system **with mediators**
  - The data collector can delegate data collection and data arrangement tasks to mediators **in a secure way** so that no mediator can know (unallowable information of) actual data.
  - Mediators can **check a data format without knowing data itself** so that data belongs to a certain range
    - age, gender, disease and so on
    - can sort out (encrypted) data by regarding a range as a quasi-identifier.
  - There is **no interaction between data suppliers and data collector**, i.e. no data supplier is required to be on-line during the data collection procedure.
  - **Ciphertext Overhead: O(1)** (in total, comm. Overhead O(n))
  - Give **formal cryptographic security definitions** (semantic security, anonymity, and format-check soundness
    - Provably secure