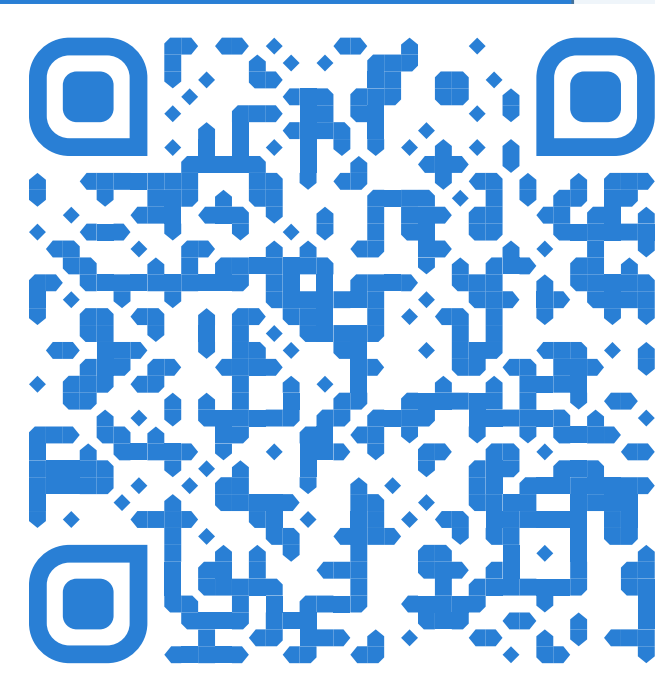


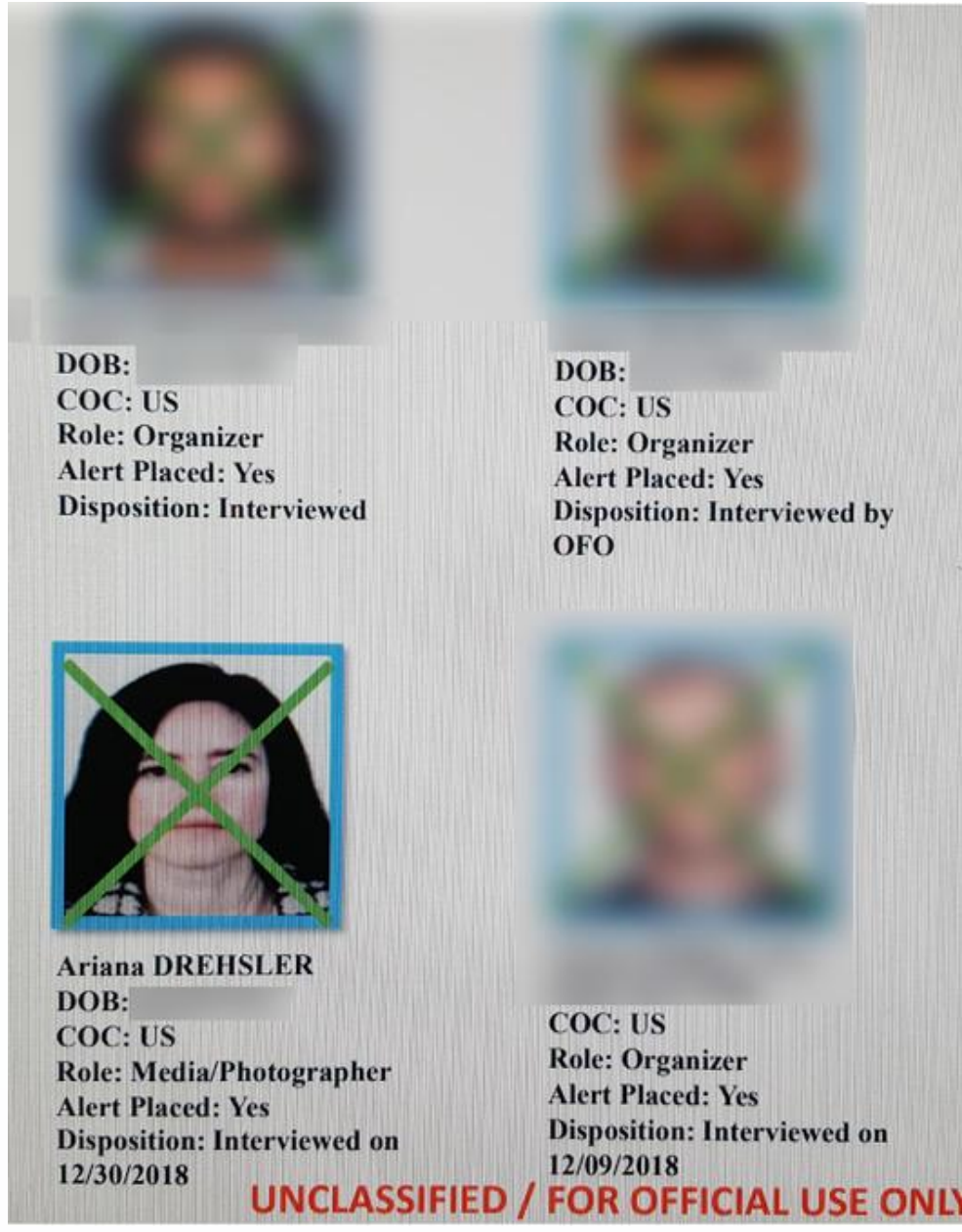
DatashareNetwork: a decentralized privacy-preserving search engine for investigative journalists

Kasra EdalatNejad (SPRING, EPFL), Wouter Lueks (SPRING, EPFL), Julien Martin, Soline Ledésert (ICIJ), Anne Lhôte (ICIJ), Bruno Thomas (ICIJ), Laurent Girod (SPRING, EPFL), Carmela Troncoso (SPRING, EPFL)



Finding relevant documents

Motivation

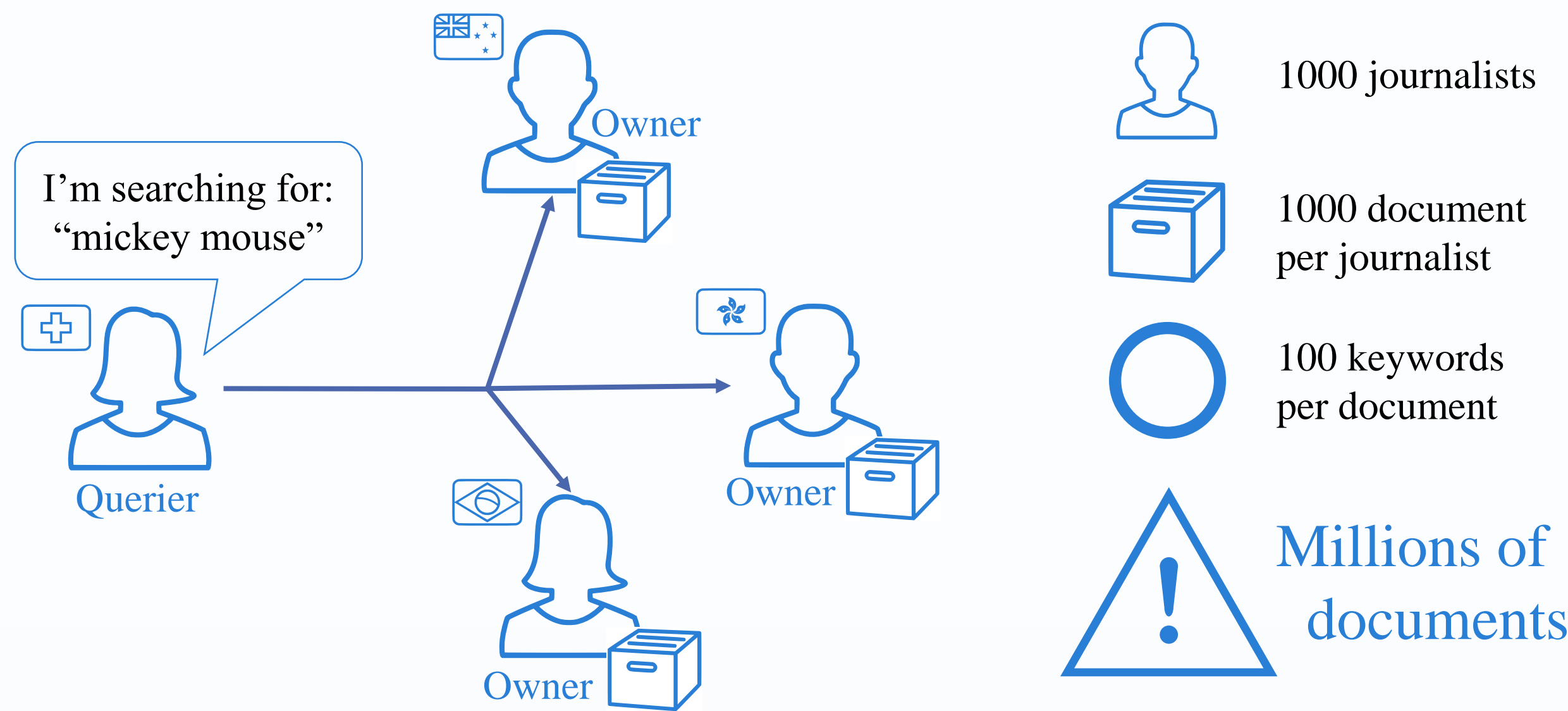


Investigative journalists gather numerous documents. These documents may reveal information about (1) what these journalists are working on and (2) the sources of these documents.

At the same time, by sharing these documents journalists gain access to more information which benefits their research. We present DatashareNetwork a system which lets journalist find and contact colleges with relevant documents.

“The leaked document shows how the US government tracks and questions journalists involved in immigration”
Source: www.nbcсандiego.com

Setting



Goals



Scalability Query privacy Document privacy Decentralized

DatashareNetwork lets journalists find and contact colleges with relevant documents

DatashareNetwork operates in two steps

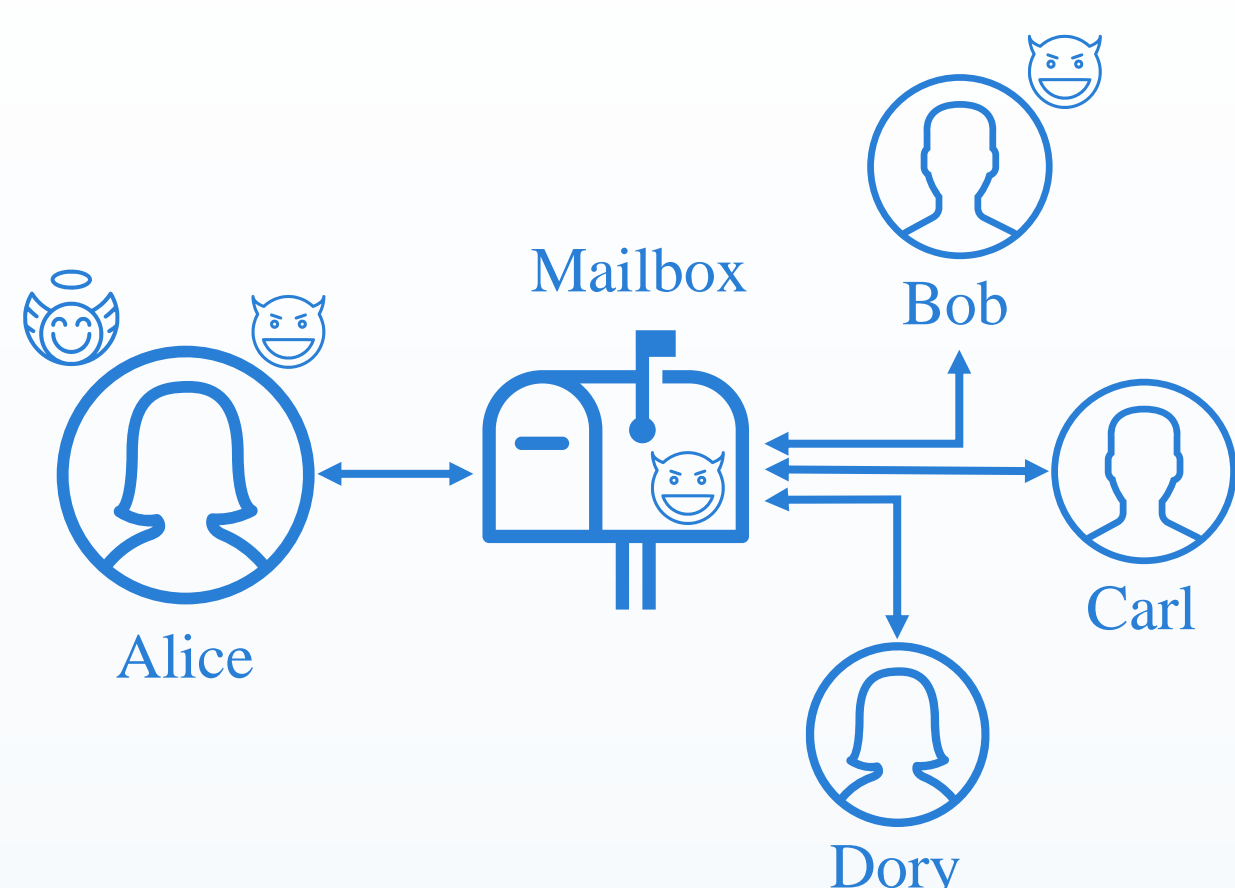


Search for documents



Contact the owner

System model



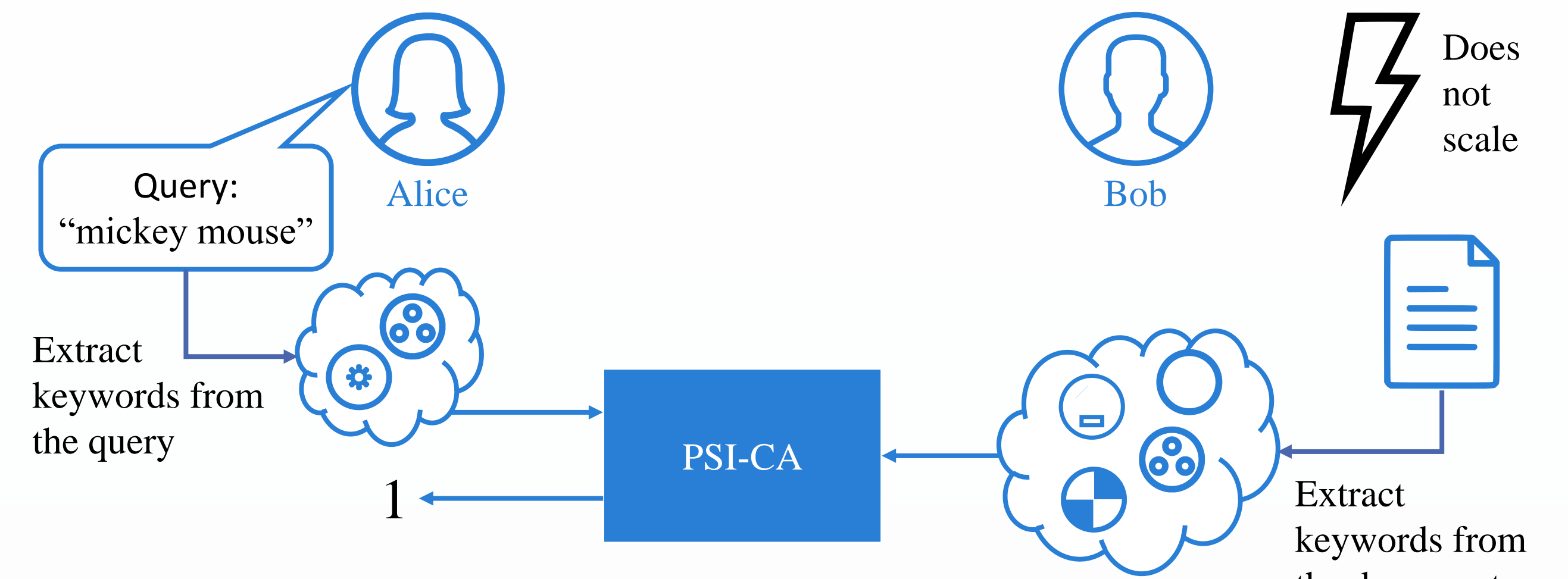
Journalists' machines can be compromised. Therefore, we cannot assume that they are honest.

- The querier may be malicious.
- The owner may be malicious.
- Mailbox is only trusted for availability.

Journalists may not be online at the same time. The mailbox store and forwards messages to facilitate the connection.

Privacy-preserving search

Existing work: private set intersection cardinality (PSI-CA)

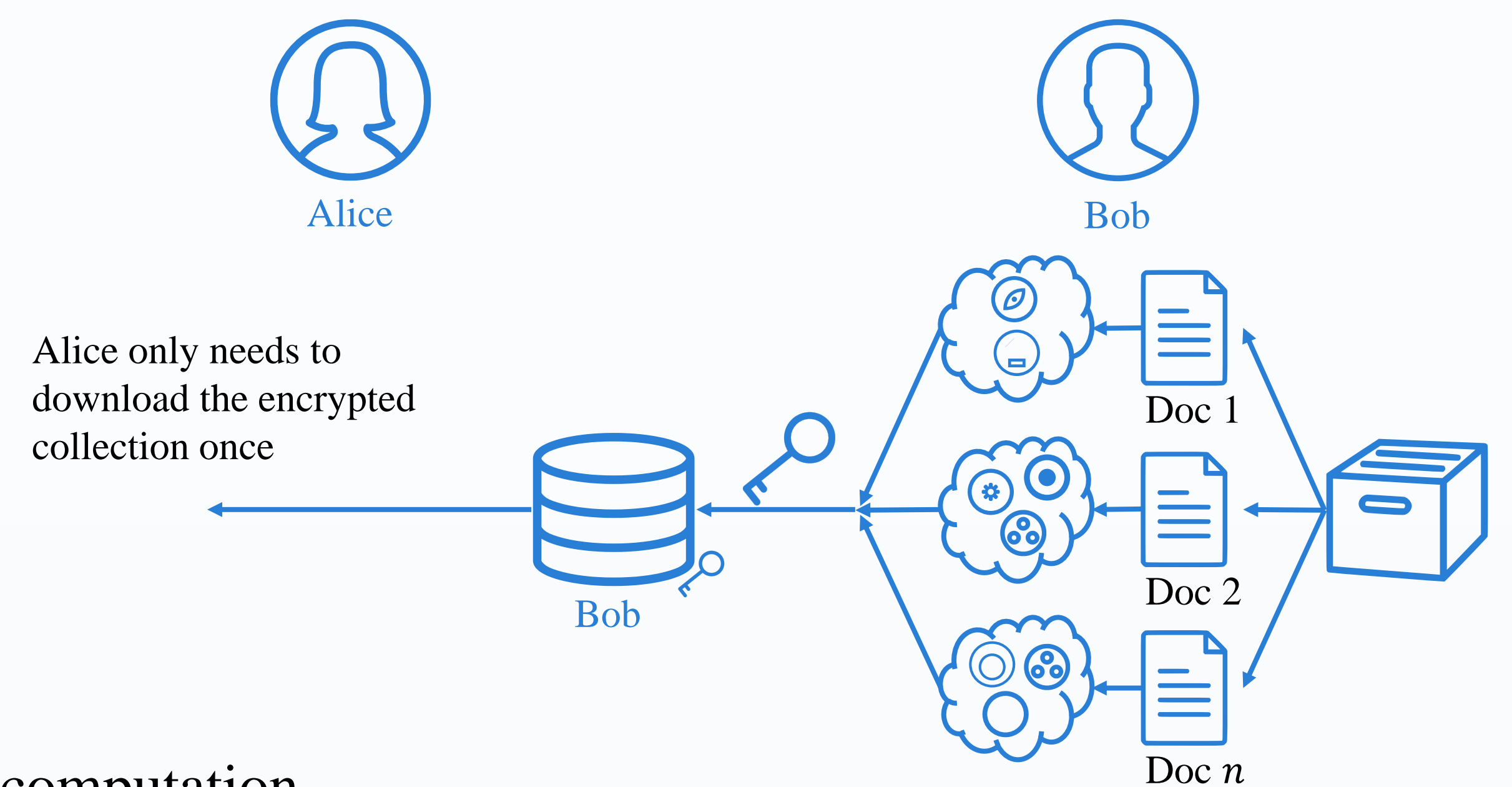


Private set intersection cardinality computes the size of intersection between two sets without revealing any information about elements. Intersection of two sets is a measure for relevance.

Multi-set private set intersection cardinality (MS-PSI-CA)

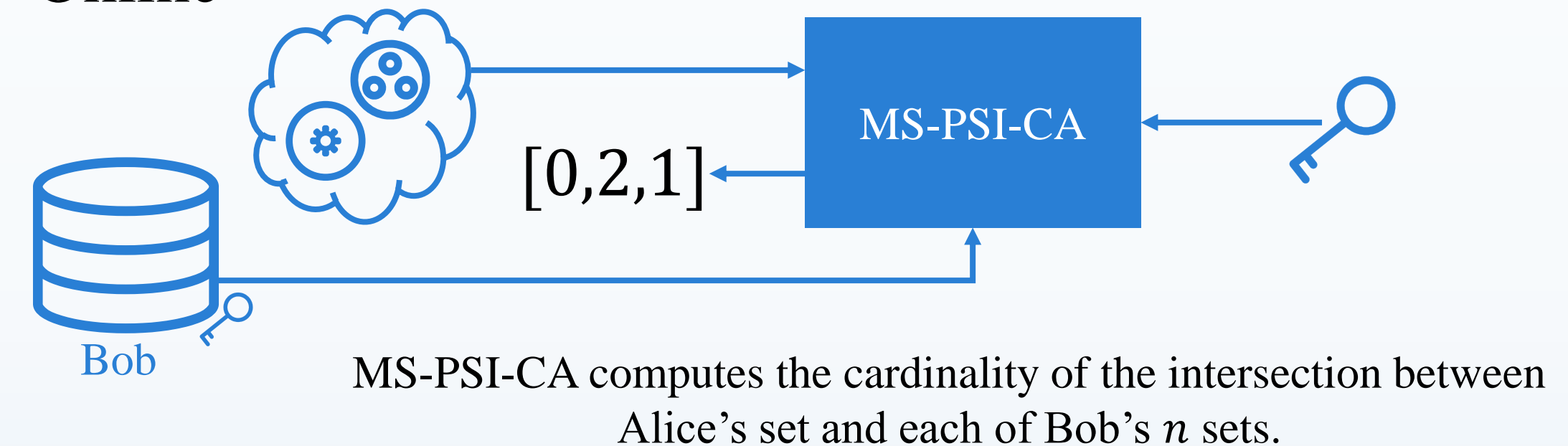
Key ideas:

Pre-compute/reuse part of the response
Query many documents at the same time



Pre-computation

Online



Evaluation

	PSI	MS-PSI-CA
Online phase		
Querier comp.	36 min	100 sec
Querier comm.	3.84 GB	640 KB
Owner comp.	12 sec	1 ms
Privacy		
Querier	✓	✓
Owner	✓	~

The table shows an estimate of the computation and communication cost of the system.

MS-PSI-CA requires a one time transfer of ~200 MB in the pre-computation.

Note that owners receive hundreds of queries per day, and the higher cost of PSI is considerable at the end of the day.

Leakage analysis

Query



Datashare doesn't leak any information about the query.

Keywords



Adversaries can always confirm the presence of keywords.

Journalist's documents



Even an ideal search protocol leaks information about documents. Datashare scales very well but does leak at a higher rate.