Promote Information Sharing by Content Hiding
(Data Anonymization)

Lukas Kencl (IRC), Gianluca Iannaccone (IRC)
Martin Loebl (CU Prague), Jose Zamora (UChile)
Tanzeem Choudhury (IRS)
Rahul Sukthankar (IRP)

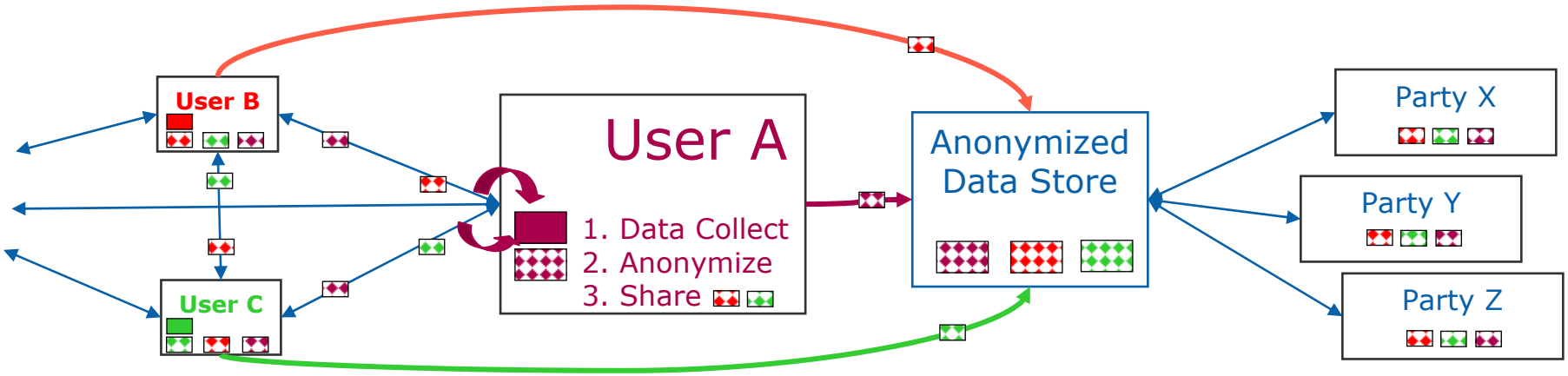
Network content hiding - motivation

- Can we share network traces? CryptoPan [ICNP'02] allows for sharing with anonymized addresses.
- With payloads?
- How to protect private content, but preserve useful context?
  - Compromise between encryption and plain data sharing
  - Algorithmic techniques to separate content & context
- Proliferation of data collection devices, privacy disappearing
  - scale
  - insider vs. outsider protection
  - some data mining useful
  - e.g. recent AOL searches trace release

Need for content-hiding, context preserving techniques

Promote Information Exchange and Shared Analysis among Untrusted Parties, while Hiding Private Content.

- protect users' private content, with computational guarantees
- allow restricted analysis and mining of shared data
- general techniques, but tailored for domain-specific knowledge



Why? Applications! Scenarios:

- Network Content Anonymization
  - Share network traces with packet payloads, enable home troubleshooting or malicious content detection (e.g. worms).
  - Online behaviour shared analysis: efficiency, self-improvement.
- Voice Anonymization, Image/Video Anonymization
- Medical, Biological, Sensor Data, …

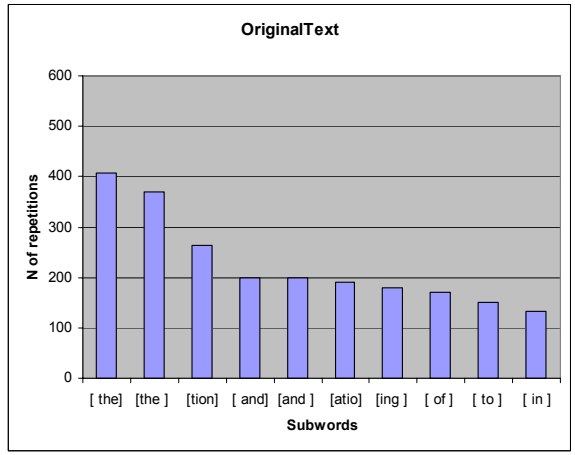
Example Prototype Technique:

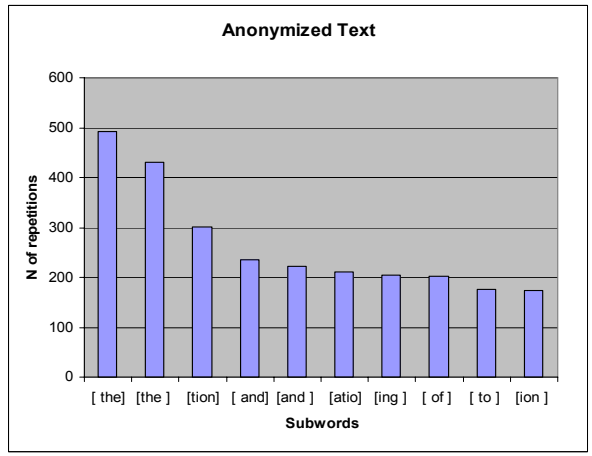
- Good Parties obtain: all substrings of length **k** (allow pattern matching, statistical analysis).
- Malicious Parties: the same, but computationally hard to reconstruct longer segments.
- Domain Knowledge: extra local treatment for sensitive segments.

Properties

- Privacy - Protecting Content Semantics
  - Individual content block reconstruction is provably Computationally Hard
- Enabling Data Analysis
  - Pattern Matching over short content blocks
  - Frequency Statistics hold over aggregates data
- Other prior art (DB community):
  - masking, scrambling, perturbation
  - usefulness depends on domain and content&context definition

Example: Pattern Frequency Ranking Holds





UN Charter text, size 32940, patterns length 4 preserved, anonymized by blocks of size 1500.

Radical change of data ownership paradigm

- Anonymized data shared collectively.
- Computational privacy guarantees promote user information exchange.
- Share with a larger community.

- *Why?*
  - Increase information exchange
  - Novel applications operating on the shared data
  - Some very cool math problems

- *Issues*
  - Techniques
  - Deployments

The End – Thank You!

Comments? Suggestions? Pointers?

Want to share traces?

<lukas.kencl@intel.com>

Backup