Aalto University
School of Science

ANTA Thematic Summer on Private Information Retrieval and Distributed Data Storage

Aalto University, Department of Mathematics and Systems Analysis

Introduction

The Algebra, Number Theory, and Applications (ANTA) research group launches a thematic summer concentrating around various themes related to private information retrieval, distributed data storage and their interplay. To this end, the group will host a visiting professor Salim El Rouayheb alongside his PhD students, as well as a number of ANTA BSc students and Aalto Science Institute (AscI) summer interns working on related topics. Throughout the summer, an intensive course, several talks, and a one-day workshop will be organized.

The course, talks, and the workshop will be announced on the ANTA seminar webpage:

`https://math.aalto.fi/en/research/discrete/anta/seminar.php`

You can subscribe to seminar announcements here:

`https://list.aalto.fi/mailman/listinfo/anta-seminar`

Welcome to attend the events and to discuss with us!

Distributed Data Storage

In modern data centers, data is stored over thousands of servers. On a global scale, large actors such as Google, Facebook, Microsoft, Amazon and Apple have a server park in the order of magnitude of a million servers, and the total amount of data stored worldwide is measured in zettabytes. This poses challenges in terms of physical storage space, energy consumption, bandwidth and security.

A key task of any data storage system is to protect the stored files even in the case of a temporary or permanent server failure. An obvious way to do this is to replicate the data, storing the same file over several nodes, thereby being able to reconstruct a back-up version of the file from a secondary server. While this simple scheme is still applied by many big actors, it is very wasteful in terms of storage space, and easy to improve on as the simple example in Figure 1 shows.

Important invariants to measure the behaviour of storage codes are the number n of servers, the number k of data items stored, the size q of the alphabet, and the largest number $d-1$ of erasures that can be tolerated. Classical trade-offs between these invariants include the Singleton bound and the Gilbert-Varshamov bound.

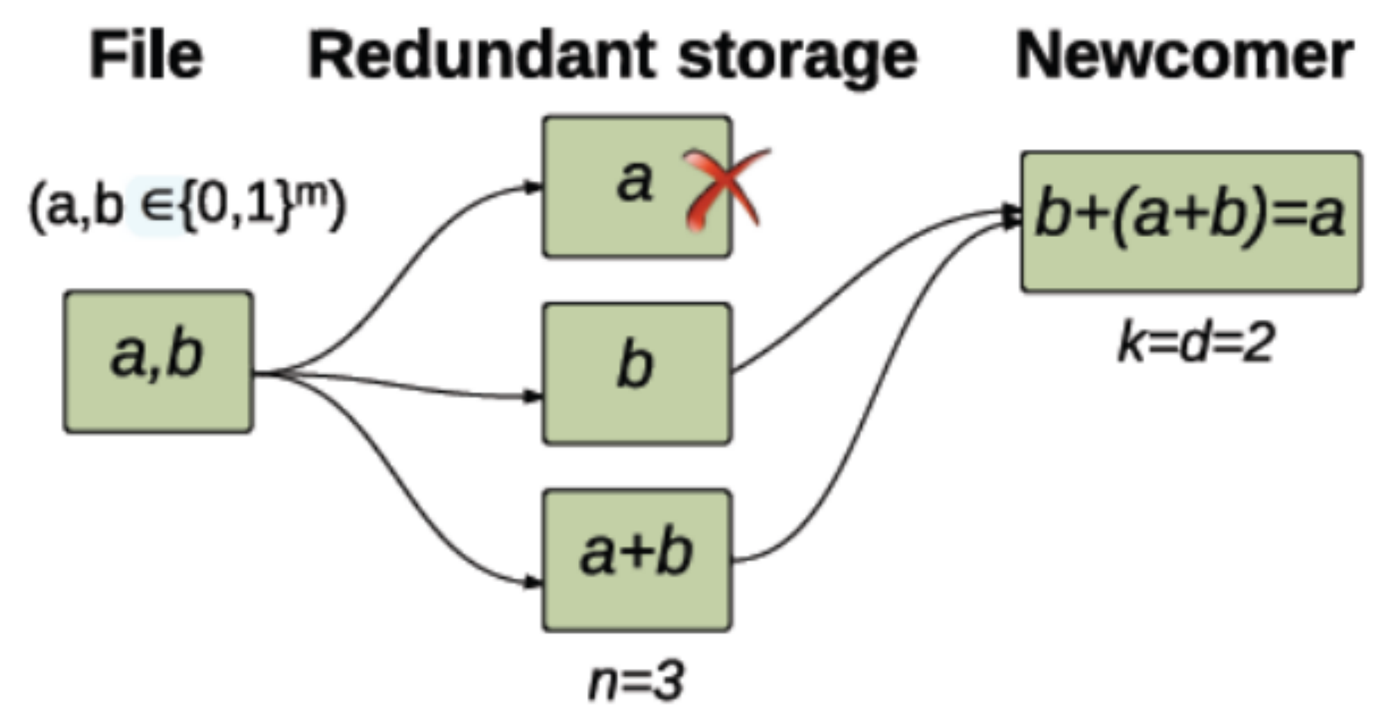


Figure 1: Storing $k=2$ files over $n=3$ nodes, tolerating $d-1=1$ failures.

Private Information Retrieval

Private information retrieval protocols make it possible for users to retrieve data items from a database without disclosing information about the identity of the data items retrieved. The notion of private information retrieval (PIR) was introduced by Chor, Goldreich, Kushilevitz and Sudan in [1] and [2]. The classic PIR model of [2] views the database as an n-bit binary string x and assumes that the user wants to retrieve a single bit x_i without revealing any information about the index i. The trivial solution is to download the entire database. This, however, incurs a significant communication overhead whenever the database is large, and is therefore not useful in practice. Unfortunately, Chor *et al.* showed in [2] that, for *information-theoretic privacy*, the trivial solution is the best one in the case of a single database stored on a single server.

This can be remedied by replicating the database onto k servers that do not communicate. This is known as k-server PIR and it has been shown ([3], [4])that is is possible to achieve sub-polynomial communication complexity.

We will now give an example of a simple 2-server PIR scheme. We have two servers, S_1 and S_2, which both store the entire n-bit database x, and Alice wants to retrieve the i-th bit x_i, for some $i \in \{1,\ldots,n\}$. To do this, Alice selects $a \in \mathbb{F}_2^n$ uniformly at random and sends a and $a+e_i$ to S_1 and S_2 respectively. The servers respond with $a \cdot x$ and $(a+e_i)\cdot x$ respectively. Given these responses Alice can compute the i-th bit as

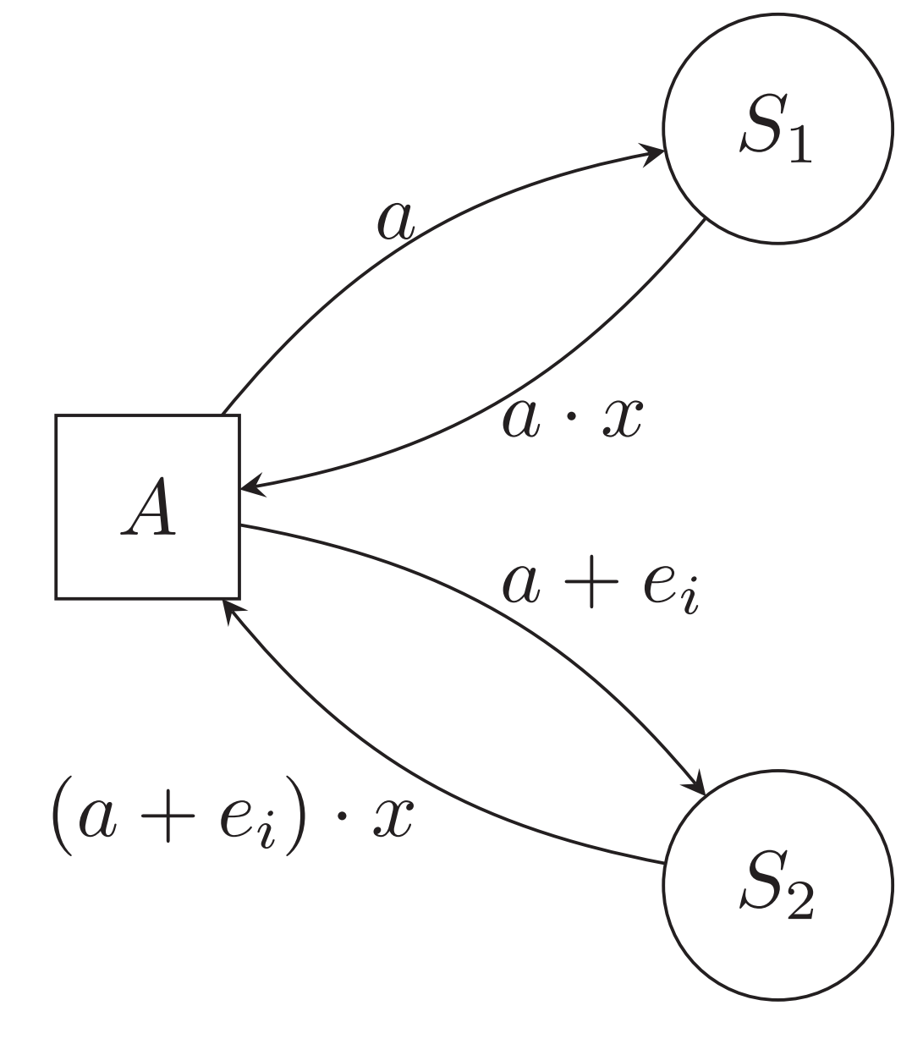
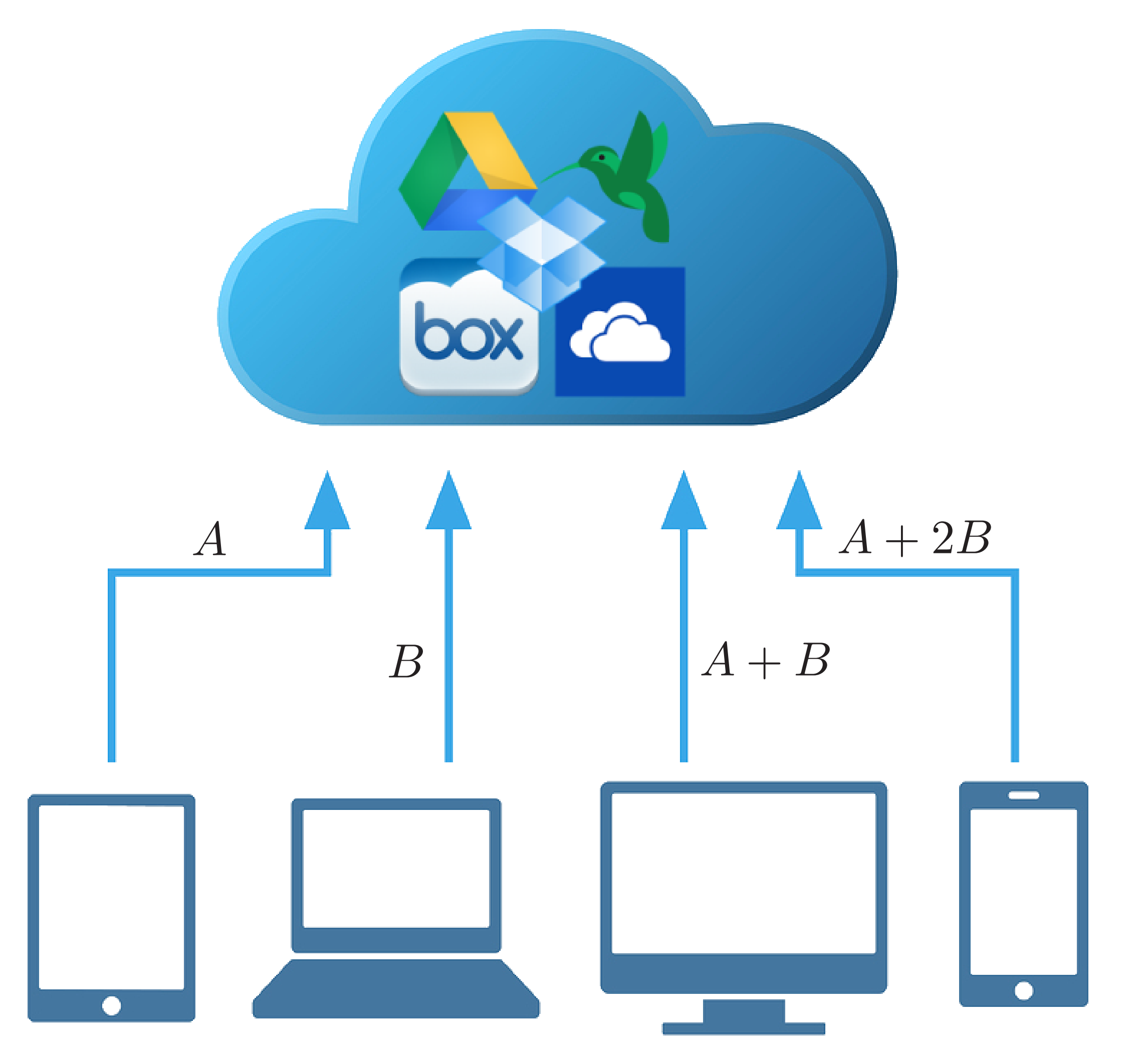


Figure 2: A simple 2-server PIR protocol.

$$a \cdot x + (a+e_i)\cdot x = (a+a+e_i)\cdot x = x_i.$$

This scheme is depicted in Figure 2. Assuming that S_1 and S_2 do not communicate, we note that the value of i remains private since a is chosen uniformly at random.

Intensive course on Coding Theory for Distributed Data Storage



Course Title: Coding Theory for Distributed Data Storage (3 cr)

Instructor: Prof. Salim El Rouayheb, Illinois Institute of Technology, Chicago

Course Description: Distributed storage systems are becoming a vital infrastructure of today's society by allowing to store reliably large amounts of data online in the "cloud" and make it accessible anywhere and anytime. In these systems, failure is the norm rather than the exception. And, to protect against data loss data is stored redundantly using codes. This course focuses on the recent state-of-the-art research topics in coding theory for distributed data storage systems. The topics covered by the course include regenerating codes, locally repairable codes, index codes, information theoretic tradeoffs and information theoretic security and privacy. Moreover, the course will also highlight how these theoretical results are currently being applied in real-world systems, such as Microsoft and Facebook data centers.

Credits: If you give a presentation and attend all 12 hours (some exceptions can be accepted), you can earn 3 credit points for this course.

Prerequisite: Some knowledge in coding theory will be helpful. Ask about participation from Salim or Camilla, if you are unsure! We can give you some reading before the course if needed.

Course Text: Lecture notes provided by the instructor and a collection of recent research papers on the topic (will be specified in class)

Organization: The course will consist of 12 hours and will be divided into two parts. The first part will consist of lectures by the instructor. The second part students will present research papers in class.

Time and place: The course will take place at Aalto during the week 30.5.-3.6. The exact time slots will be decided together with the students.

Registration: E-mail Camilla by 15.5. Tell shortly about your background (relevant courses taken and why you are interested) and (un)preferred time slots, if any.

References

[1] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 41–50. IEEE, 1995.

[2] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private information retrieval. *Journal of the ACM (JACM)*, 45(6):965–981, 1998.

[3] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 39–44. ACM, 2009.

[4] Zeev Dvir and Sivakanth Gopi. 2-server pir with sub-polynomial communication. *arXiv preprint arXiv:1407.6692*, 2014.

Contact Information

Camilla Hollanti, Associate Professor
Department of Mathematics and Systems Analysis
Room Y239, Otakaari 1, Espoo
Aalto University

Web `https://math.aalto.fi/en/people/camilla.hollanti`
Email camilla.hollanti@aalto.fi
Phone +358 50 562 8987

Salim El Rouayheb, Professor
Illinois Institute of Technology, Chicago

Web `http://www.ece.iit.edu/~salim/`
Email salim@iit.edu