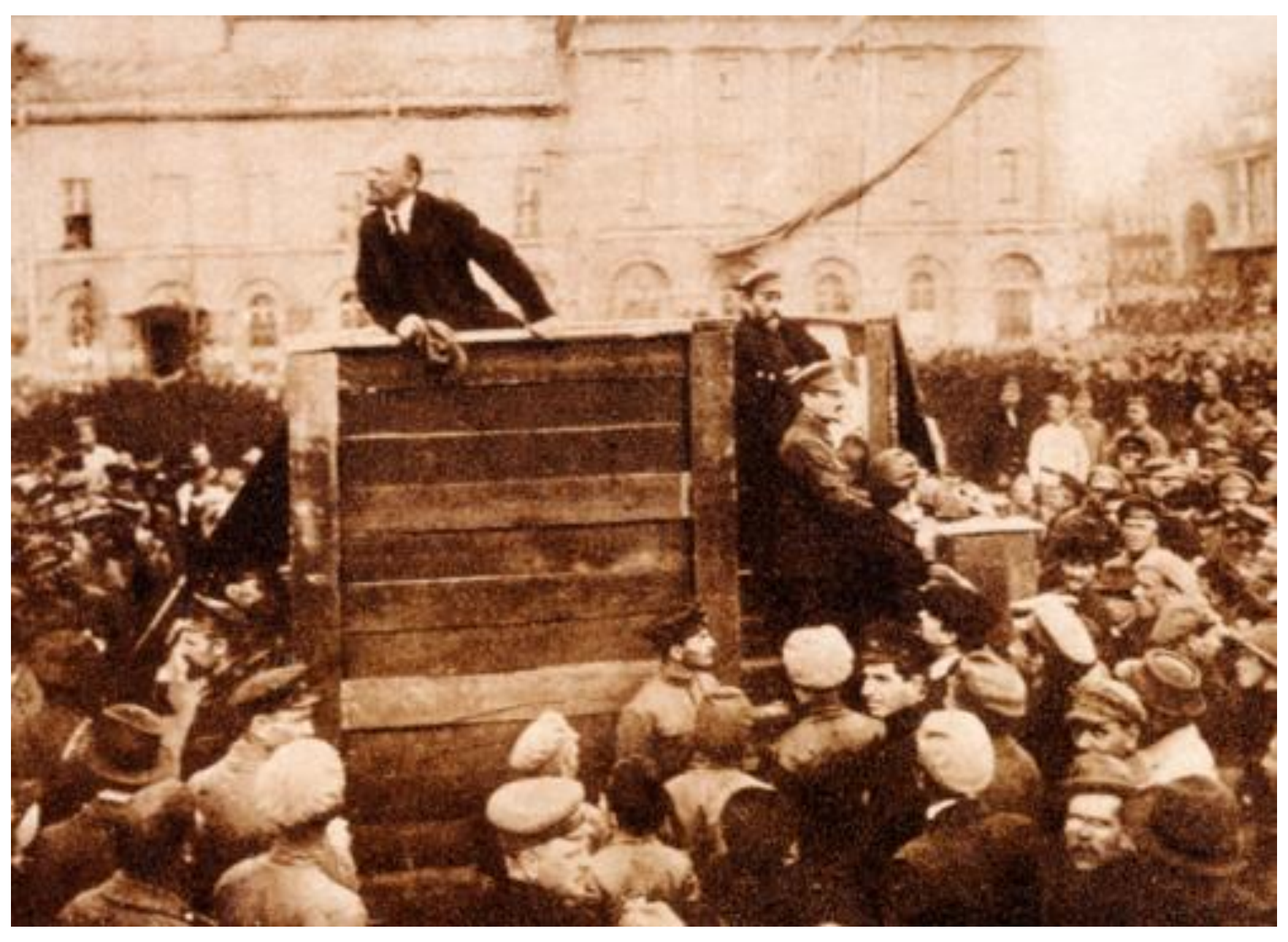
Informatics

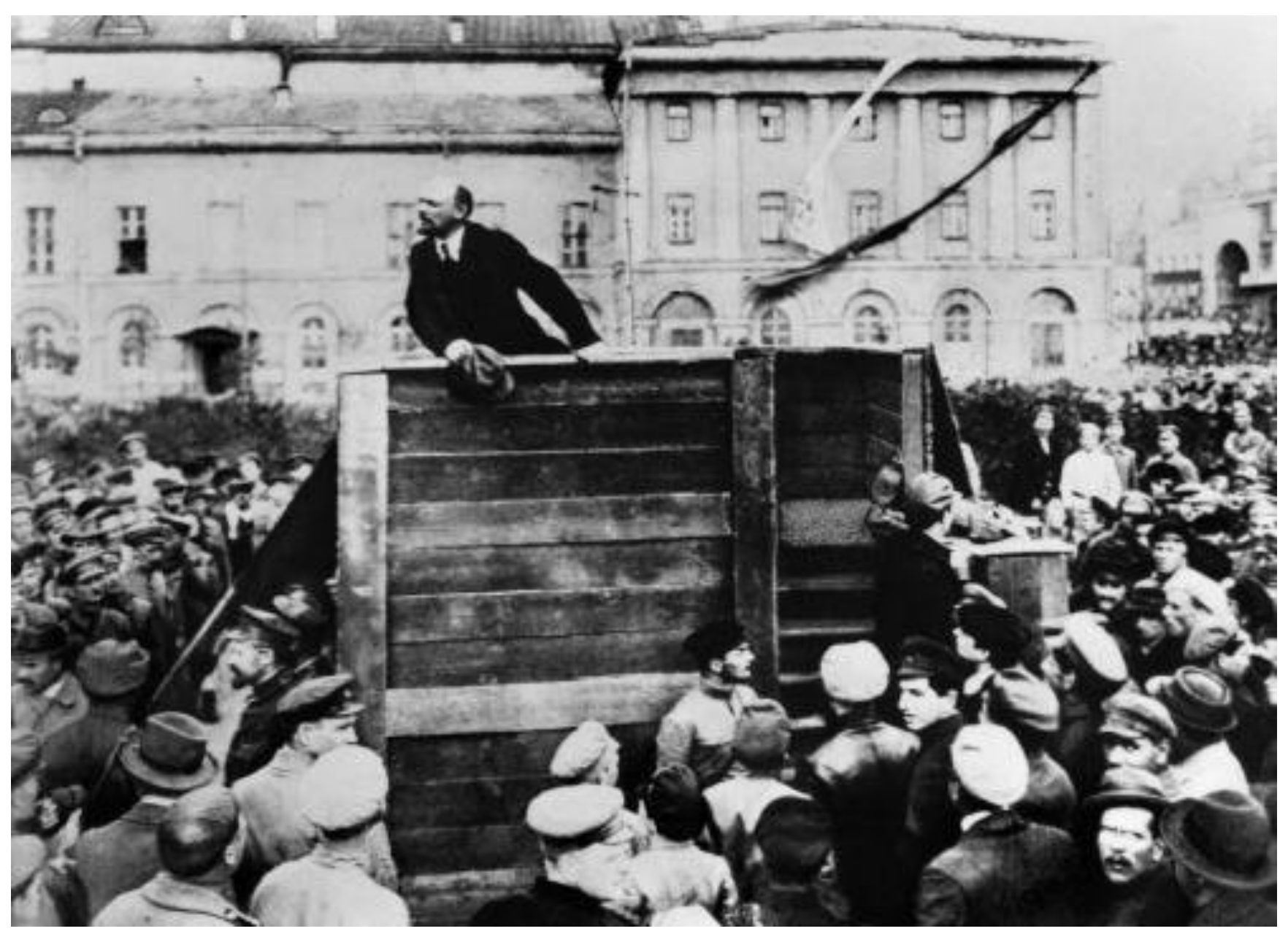
Reducing the operation cost of a file fixtiy storage on the ethereum blockchain by utilizing pool testing strategies

Michael Etschbacher
Data Science

TU Wien Informatics
Institute of Information Systems Engineering E194
Information and Software Engineering E194-01
Supervisor: Ao.univ.Prof. Dr. Andreas Rauber
Contact: etsch.michael@gmail.com

Motivation

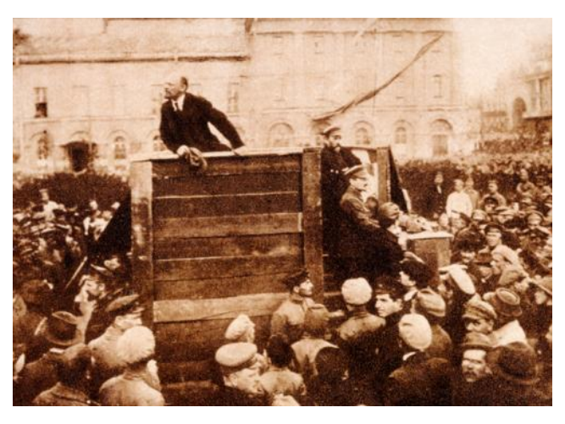




Storing cultural heritage in digital archives offers malicious actors the possibility to manipulate the data and possibly forge history. Recent digital technologies make data manipulation more efficient andless costly. In 1920 a photography was taken of Vladimir Lenin atop a platformspeaking to a crowd. In the original photo, Lenin's comrade Leon Trotsky can be seen standing beside the platform on Lenin's left side. When power struggles within the revolution forced Trotsky out of the party 7 years later, he was retouched out of the picture, using paint, razors and airbrushes [HS05].

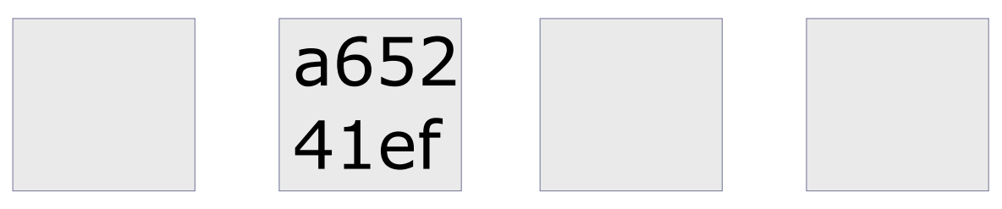
How to detect data manipulation

1 Calculate a cryptographic hash of your object, so-called fixity information.
SHA256 is utilized in this thesis



a65241efd9d82
ae28cb1d04d3b
1aaf206bd8d71
90ce241ca6a59
3f22a994678c

2 Store the value securly, considering that a malicious actor that is able to alter the fixity information can also alter the underlying object illicit.
The Ethereum blockchain is used as a immutable storage in this thesis



3 On retrieval of your object from the archive, calculate a hash value of the retrieved object.
Compare the persisted hash on the blockchain with the newly calculated one. If the hashes match, the object is guaranteed to be unchanged [KGA+14].

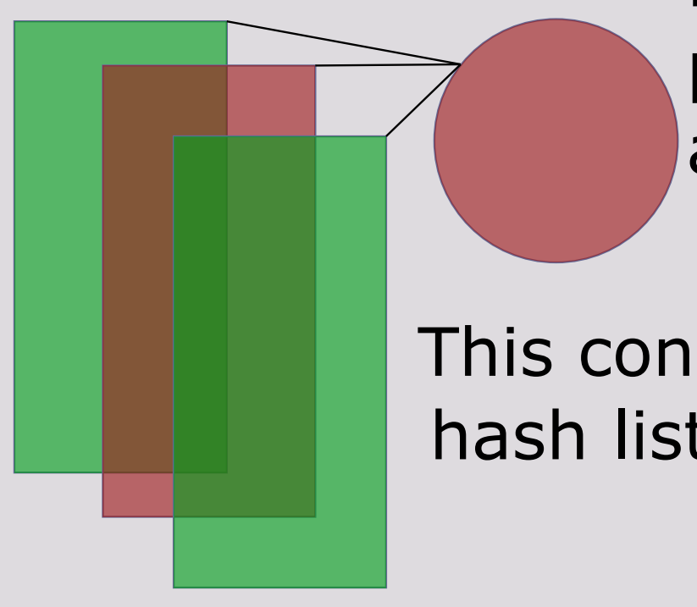
d4565aff917f74
b9009d6998ff3
b4689c68b4f9d
f7eba5c4a1a6a
28abe4d14b1

a65241ef... != d4565aff...

Problem

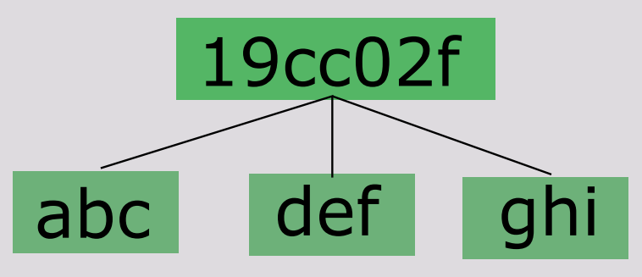
The cost of storing a cryptographic hash value on the ethereum blockchain **foreach** digital object in the ingestion process is too high

Solution



Pooled testing, established during the COVID-19 pandemic, where multiple probes are combined and only the pool has to be tested [ŽLG21]

This concept can also be implemented with hash lists, where only the root hash has to be persisted on the blockchain.



Relevance Cycle

- Ethereum blockchain functions as a immutable distributed database to store fixity information [Dan17] [WY21]
- Project ARCHANGEL has shown that the Ethereum blockchain is suited to store fixity information [CBB+18]
- The cost of storing the SHA256 value of a digital object on the blockchain costs about **$5**
- The artefact must reduce the amount of costly write transaction, by utilizing pooled testing, by at least **50%**

Design Science Research
by Hevner [Hev07]

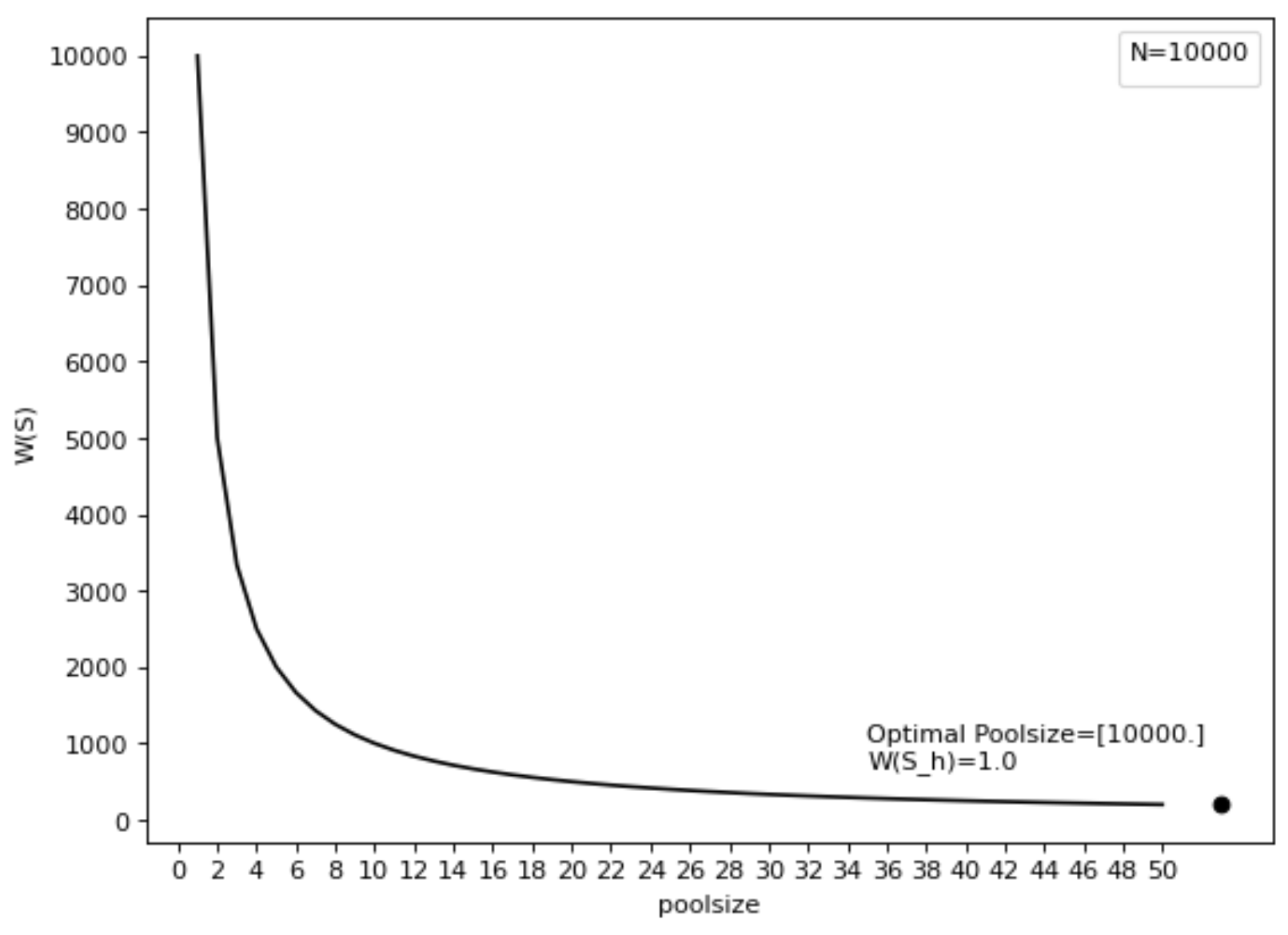
Design Cycle

- Implementation of the fixity storage service, as a Smart Contract written in Solidity.
- The storage service must expose CREATE,UPDATE and READ methods to manage fixity information.
- Implementation of two pooling strategies: (1) two-stage-hierarchical pooling and (2) context-sensitive pooling [DBK20]
- Evaluation of the presented strategies compared to an individual testing strategy in terms of efficiency and operation cost

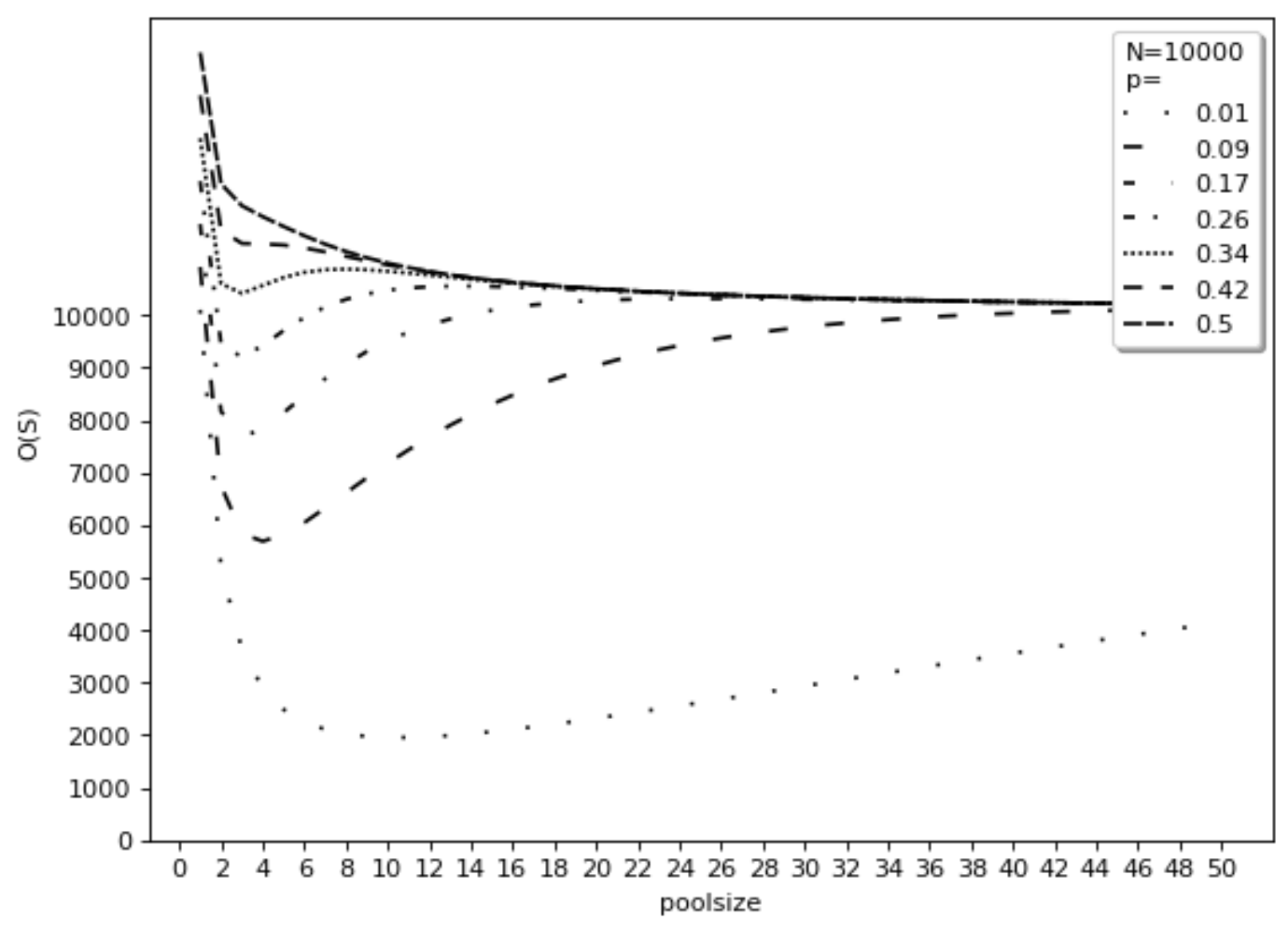
Rigor Cycle

- Collomosse et. al (2018) proposed ARCHANGEL, which is a file fixity storage service on a private fork of the Ethereum blockchain [CBB+18]
- Sigwarts et. al (2020) utilized the Ethereum blockchain as a storage for provenance data [SBP+20]
- Pooled Testing was first introduced by Dorfman as a strategy to screen a large amount of recruits for syphilis during World War 2 [Dor43]
- Deckert et. al (2020) analysed a context-sensitive pooling pooling approach for COVID-19 mass testing [DBK20]

RQ 1 What is the optimal pool size based on the change rates of digital objects in the archive regarding cost and efficiency?



x: pool size, y:#blockchain-writes

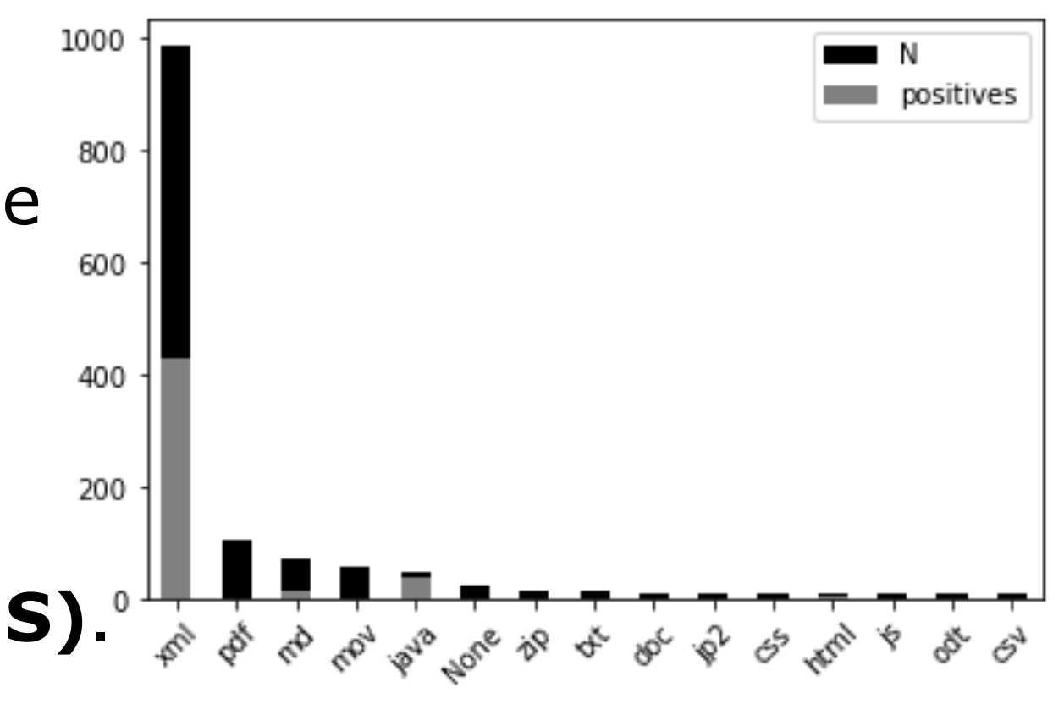


x: pool size, y:#total operations needed

The optimal pool size, regarding the cost is **N**, as seen in the left figure above, considering that if you only have one pool, you only have to write once to the blockchain.
But, if you want to retrieve an object from your archive, you have to recompute the large pool, and if the pool is corrupted you have to replace N objects with correct copies, so called data-scrubbing.
Counting in the data-scrubbing operations, optimal pool sizes from **2 to 10** are favorable, as seen in the right figure above

RQ 2 To what extent can pooled testing increase the efficiency and reduce cost for a fixity information storage service on the Ethereum blockchain?

The *https://github.com/openpreserve/format-corpus* is an openly-licensed corpus, consisting of 1560 files of various formats and creation tools. The dataset has an overall change rate of 32%, meaning that almost every third file has experienced at least one change over the course of storage. The total #operations needed **O(S)**, in order to preserve the dataset consists of the #writing operations **W(S)** on the blockchain added to the local #data-scrubbing operations **R(S)**.



Pooling-Strategy S	Operations O(S)	Efficiency E(S)	Writes W(S)	Cost	Cost Efficiency C(S)
Individual S_i	2070	1.00	1560	$7,800	1.00
Two-Stage-Hierarchical S_h	1605	1.28	520	$2,600	3.00
Context-Sensitive S_{cs}	1491	1.38	662	$3,310	2.35

RQ3 Given that metadata has a higher change rate, what effect has the split of metadata and objects on the operation cost?

The idea behind this approach was to create a group of highly volatile files, the metadata and a group of stable digital objects. The resulting dataset, where the metadata is split-off the object consists of the double amount of files, 3120. The large pool sizes in the non-volatile group could not compensate for the double amount of files overall.
The result, shown in the Table beside, that this approach is worse in terms of efficiency and cost than the approaches with non-split off metadata

p	p_{meta}	$E(S_{cs})$	$C(S_{cs})$	$E(S_h)$	$C(S_h)$
0,001	0,99	0,77	1,71	0,86	1,91
0,007	0,96	0,73	1,60	0,80	1,76
0,013	0,94	0,71	1,55	0,78	1,69
0,038	0,84	0,67	1,46	0,73	1,58
0,044	0,81	0,66	1,42	0,71	1,52

[HS05] Mark Hofer and Kathleen Owings Swan. Digital image manipulation: A compelling means to engage students in discussion of point of view and perspective. Contemporary Issues in Technology and Teacher Education 5(3):290–299, 2005.
[Hev07] Alan R Hevner. A three cycle view of design science research. Scandinavian journal of information systems, 19(2):4, 2007.
[Dor43] Robert Dorfman. The detection of defective members of large populations. The Annals of Mathematical Statistics, 14(4):436–440, 1943.
[ŽLG21] Julius Žilinskas, Algirdas Lančinskas, and Mario R Guarracino. Pooled testing with replication as a mass testing strategy for the COVID-19 pandemics. Scientific Reports, 11(1):1–7, 2021.
[DBK20] Andreas Deckert, Till Bärnighausen, and Nicholas NA Kyei. Simulation of pooled-sample analysis strategies for COVID-19 mass testing. Bulletin of the World Health Organization, 98(9):590, 2020.
[CBB+18] John Collomosse, Tu Bui, Alan Brown, John Sheridan, Alex Green, Mark Bell, Jamie Fawcett, Jez Higgins, and Olivier Thereaux. Archangel: Trusted archives of digital public documents. In Proceedings of the ACM Symposium on Document Engineering 2018, pages 1–4, 2018.
[KGA+14] K. Kim, W. Graham, D.S. Alliance, A Reich, and C. Kussmann. What is Fixity, and When Should I be Checking It. https://osf.io/an5zh/, 2014.
[SBP+20] Marten Sigwart, Michael Borkowski, Marco Peise, Stefan Schulte, and Stefan Tai. A secure and extensible blockchain-based data provenance framework for the Internet of Things. Personal and Ubiquitous Computing, 2020.
[WY21] Hosung Wang and Dongmin Yang. Research and development of blockchain recordkeeping at the National Archives of Korea. Computers, 10(8):90, 2021.
[Dan17] Chris Dannen. Introducing Ethereum and solidity, volume 1. Springer, 2017.