September 2019

114Similar Systems Found in Lawsection.4

Cryptographic Methods with a *Pli Cacheté*

Towards the Computational Assurance of Integrity

Thatcher L COLLINS [a,1]
[a] *Department of Applied Mathematics, University of Washington, Seattle, USA*

Abstract. Unreproducibility stemming from a loss of data integrity can be prevented with hash functions, secure sketches, and Benford's Law when combined with the historical practice of a *Pli Cacheté* where scientific discoveries were archived with a 3rd party to later prove the date of discovery. Including the distinct systems of preregistation and data provenance tracking becomes the starting point for the creation of a complete ontology of scientific documentation. The ultimate goals in such a system–ideally mandated–would rule out several forms of dishonesty, catch computational and database errors, catch honest mistakes, and allow for automated data audits of large collaborative open science projects.

Keywords. reproducibility, hash function, secure sketch, fuzzy extractor, data fraud, library science, open science, applied ontology, publication bias, plagiarism, data provenance, preregistration, Benford's Law

1. Introduction

When integrity breaks down in a scientific setting, the mess can involve legal action, investigations, accusations, and negative media coverage. To prevent that, a systematic and unbiased way to prevent fraud or inadvertent corruption of the data or the results is proposed. The *European Code of Conduct for Research Integrity* defines integrity as "**Reliability** in ensuring the quality of research, reflected in the design, the methodology, the analysis and the use of resources. **Honesty** in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair, full and unbiased way. **Respect** for colleagues, research participants, society, ecosystems, cultural heritage and the environment. **Accountability** for the research from idea to publication, for its management and organisation, for training, supervision and mentoring, and for its wider impacts"[1]. Except for respect, training and mentoring, a loss of integrity can lead to unreproducible science. Within that subset of integrity issues, the hardest part is finding the resultant hidden changes. Thus the practical manifestation of integrity in data-driven science is: **no unintentional changes, no secret changes**. That is, every significant change in content is intentional and tracked over time and space; data provenance expands to include provenance for the application of the scientific method.

[1] E-mail: *thtchr@uw.edu.*; https://orcid.org/0000-0003-0591-0823; Project Page: https://osf.io/asbtg/

Fraud includes any human misconduct that changes the science (most often for data but not necessarily), including plagiarism, manipulation, or fabrication. For this paper, **corruption** always refers to the sort with data, not political corruption which is simply a motivation or qualitative description of the machinations of an act of fraud. In a broad-form definition for what most people call plagiarism, The Council of Science Editors (CSE) defines **piracy** as "the unauthorized reproduction or use of ideas, data, or methods from others without adequate permission or acknowledgment," even by secretly repurposing one's own work. A similarly broad definition was described in the 13th century by Persian Poet Shams-e-Qays with a useful categorization of **plagiarism**: 1) Verbatim (*Intihal*): exact copy; 2) "Flayed" (*Salkh*): changing the order, re-arranging; 3) Conceptual (*Elmam*): "approaching" the exact copy in concept; and 4) Domain Transfer (*Naql*): unattributed reuse in a new domain [2]. The CSE's technical definition of plagiarism only encompasses verbatim and flayed plagiarism, the others fall under piracy [3]. Is there sufficient evidence of fraud to justify all of this work? In *Chemistry of Materials*, a 2019 study of chemistry articles published from 2017-2018 found that 42% of 331 retracted chemistry papers were retracted because of plagiarism and another 16% were retracted due to falsified data. Only 16% were due to honest errors [4] [5].

2. Hashing from the Start

A hash function maps a digital object to a finite uniform string called a **checksum** or a **hash** with no discernible relationship to the original object, acting as a secret identification code. Two objects that are exactly the same will always produce the same hash. Two identical hashes are likely to come from the same object, but not necessarily (because an infinite list of potential objects mapped to a finite set of strings will sometimes produce a **collision**). One change to an object will produce a completely different hash. Hashes allow for rapid checks of changes in content[6]. Hashing prevents Shams-e-Qays's verbatim-type plagiarism.

For version control such as Git and data management, hashing is already essential [7]. In a recent overview of reproducibility systems, The Whole Tale Project [8] mentioned "an optional checksum" of data. In this model, integrity checks **within** a system (like Git) are standard, but integrity checks **between** different research components are optional. If a mandatory automatic checksum is standard elsewhere, why should it be optional for science? Moreover these are **internal** integrity checks, but scientific integrity would be better ensured by including **external** 3rd party checks. After publication or at the completion of the scientific process, data and other artifacts might go to a repository, many of which create a checksum at this point to ensure the integrity of the data as it is stored long-term or moved around for other uses. CoreTrustSeal requires checksums in their certification of scientific repositories [9]. Dryad, Zenodo, 3TU.Datacentrum (4TU), and OSF all use checksums. Dryad is notable for including an audit before final acceptance into the database [11]. Libraries often include checksums in their Data Management Policies (DMPs) [11]. The Open Science Chain is designed to handle provenance and integrity after publication where data reuse is very complicated [12]. The advantages of using a simple hash algorithm (or collection of simple ones) is that they are: portable, found on every modern operating system, unlikely to become obsolete, and fast.

3. *Pli Cacheté*: Caching all the Hashing

A checksum is only a snapshot certification of content, not timing. Unimpeachable certification of existence in time requires physical possession by a 3rd party. Thomas Erren makes a compelling case to create a modern version of a 3rd party system called a *Pli Cacheté* (French for a sealed envelope) first brought to prominence by the French *Academie des Sciences* in the 1700s which accepted draft deposits of scientific work. Erren suggests reviving the *Pli Cacheté* so that researchers have "the opportunity to claim priority of sealed scientific rationale and data which may not be substantiated enough and might mislead when published too early or even erroneously" [13]. Notice that it helps the scientist by preventing a politically charged debate over the primacy of scientific discovery (accurate recognition of which is a type of integrity), but also, it encourages systematic, careful research; "reliable" in the sense of the *European Code of Conduct*.

Expand this concept to each stage of the scientific process including each day of data collection, where a simple checksum from the hash function is given to a library (rather than an academic journal as suggested by Erren) or other 3rd party such as an open repository. Instead of simply collecting the various components needed to reproduce a scientific result, an academic publisher (or peer reviewer) can hash the received contents and audit their checksum against the 3rd party checksums. Hashing a step and sending it to a library takes less time than reading this article; easy to do, but easy to forget (this is the great implementation challenge).

The full power of a cryptographic hash comes when used as a mandatory audit at every stage of the scientific stack:

1. Hashing the **hypothesis** prevents *post-hoc* storytelling or changing the hypothesis to fit the data;
2. Hashing the text of the domain-specific **data collection methods** (e.g. lab techniques) certifies that methods to be reproduced do not suffer from differences in memory; or have not changed significantly during the study;
3. Hashing the **data** as soon as it is recorded prevents manipulating the data or changing the outlier policy;
4. Hashing **additional stages** (such as noise reduction) further adds to the difficulty of fabrication and falsification;
5. Hashing the **conclusions and results** preserves the patent rights and scientific credit; and prevents errors of publishing too soon or without sufficient evidence.

The North American Scientific Integrity Consortium highlights a primary problem: "there are impediments and disadvantages of open science that must be acknowledged, including concerns with intellectual property, matters of national security, and the potential loss of confidentiality of research participants in human clinical trials" [14]. The *Pli Cacheté* system maintains privacy even as hashes of the data can ensure integrity of private or secret data. In support of this approach, an editorial in the *Journal Nature* includes "better record-keeping" in their proposed solutions [15]

4. Similar Systems Found in Law

The Paris Convention for the Protection of Industrial Property (1883) includes a provision–still in force–that an inventor has 6-12 months to file, if desired, in the other

"Contracting States," retaining the original filing date as the date of discovery. But because the clock on the monopoly period in the US does not start until the US patent application is filed, inventors can file in Europe first, then just under 12 months later in the US, granting the inventor an extra year of patent protection [17]. In 1995, US law changed to "give U.S. applicants parity with foreign applicants under the GATT Uruguay Round Agreements." Since then, US applicants can file a **provisional patent** in a sealed envelope for up to a year, providing the same year-long grace period to filers of US patents [16]. The provisional application is not opened till the full application is filed, just like the traditional *Pli Cacheté*.

Second, common law and many other legal systems have rules for the immediate acceptance of documentary evidence (**self-authenticating** (USA), self-proving (Scotland), public instruments (British Common Law), and authentic instruments (EU)) [18]. Third, carbon copies create exact copies at the time of creation for the rapid authentication of business documents. A *Pli Cacheté*, broadened to be like self-authenticating documents can help avoid or ease the resolution of litigation, bureaucratic adjudication, and messy public battles. These similar systems are summarized in Table 1.

Table 1: Summary of Similar Systems

	Pli Cacheté	Patents	Carbon Copies	Self-Authenticated
Domain	science	technology	business	law
Purpose	integrity	monopoly rights	disputes	litagation
3rd Party	library	patent office	(varies)	government

5. In the Classroom

A 1994 study found that "Eighty-nine percent of students surveyed admitted they had cheated" [19]. Without evidence, Canada's York University has a Teaching Policy which states "Academic dishonesty is a serious problem in undergraduate labs. This is partly because the culture of lab courses sometimes fosters plagiarism." Their solution is similar to the Pli Cacheté but uses Teaching Assistants as a 3rd party: "students obtain the TAs signature on all pages of their original lab notes and data, and submit those notes with their lab report" or else a "carbon copy may then be ripped out and handed to the TA before the student leaves the class" [20]. Thus, academia is already using self-authentication to solve integrity problems. Lab classes might also require submission of a pre-lab beforehand, where a pre-lab mirrors preregistration.

6. Comparison to Preregistration and Data Provenance

Preregistration (also called registered reports) is a reaction to the nonpublication of negative science, also called **publication bias** [21] or the **file drawer effect**. The Open Science Framework has guidelines for open science that state "preregistration of studies is a means of making research more discoverable even if it does not get published. Preregistration of Analysis Plans certifies the distinction between confirmatory and exploratory research" [22]. This preregistration of analysis overlaps with the caching of a study plan, but the holder of the deposit is a journal (2nd party relationship. Preregistration, at a minimum, acknowledges the existence of negative results. At its best, it comes

with a commitment to publish regardless of the result. A call-to-action from the 2019 Hong Kong Conference for Research Integrity states "Value the reporting of all research, regardless of the results" [23]. In the event of a conflict of interest with the journal, a 3rd party *Pli Cacheté* preserves integrity.

One subtle integrity problem, suggested by Klein et al, is that "while embargoes on preregistrations can mitigate the fear of being scooped, flexibility in the release of pre-registered documents limits transparency. For example, researchers may strategically release only those documents that fit the narrative they wish to convey once the results are in. It is therefore preferable to encourage transparency from the outset" [21]. But a hash of the preregistered documents would still allow reviewers and the public to verify the completeness of an embargoed preregistration.

Data provenance is "the derivation history of a data product starting from its original sources" as well as the information needed to process, identify, and distinguish a dataset [24]. Both provenance and *Pli Cacheté* indicate the time and place of the origin of a data product; *Pli Cacheté* authenticates the time, while provenance describes the details of the journey (helpful to an auditor for investigating fraud). *Pli Cacheté* hashes the contents while provenance describes the contents. Data provenance within a project might be handled by Git; data provenance after a project might be handled by a separate metadata file or a scientific data blockchain such as Open Science Chain (OSC)[12]. Currently, Git is for tracking the creation of something while OSC is for tracking reuse and modification **after** creation. But a *Pli Cacheté* would use universally-available hash already designed for rapid match checks, allowing an integrity check in any other system.

Are all three systems necessary? Consider the standard questions in Table 2 used by journalists to ensure an accurate report: who, what, where, why, and how. Outside of that standard list is the question of corroboration: does an independent source, a second reporter, or some other previously unknown document corroborate the answers to the standard list? Compare the coverage of these questions by the three systems:

Table 2: The Three *P*s of Scientific Documentation

Question	Pli Cachete	Preregistration	Provenance
who	x	x	x
what	x	x	x
where			x
when	x		x
why		x	
how		x	
corroboration	x		
mid-stream	x		x
ownership	3rd party	2nd party	1st party

Each system's information is held by a different party **during** the scientific process, with every question answered somewhere. The goal is to use this completeness to solve reproducibility problems and aid integrity investigations.

7. Tackling the File Drawer Effect

Scargle defines publication bias and the file drawer effect to be when the probability of being published "depends on the statistical significance of its results" [25]. In effect,

amalgamated research (including metaresearch and multi-grid systems) become biased by an unrepresentative sample. Clinical studies often require preregistration by law (with no guarantee of publication). In *Selective Publication of Antidepressant Trials and its Influence on Apparent Efficacy* [26], not only is the existence and statistical effect of publication bias demonstrated, but also acknowledged is the switching of domains after a **secondary effect** shows positive results. Whether or not an amalgamated study using secondary results is biased is up to others to evaluate, but data provenance and preregistration give researchers the investigative information they need. If by fraud, a study is changed (or fabricated) to be positive, this moves a study from negative to positive. In Table 3, the three integrity systems illuminate how to prevent the three file drawer effects.

Table 3: The Filedrawer Effect

	published	not published	domain change	fraud
relationship to claim	positive	negative	positive	positive
solution	–	preregistrtion	provenance	pli cachete

8. Conflict Resolution Amongst Scientists

A dispute can occur within a scientific group during the scientific process or even much later, including during a retraction process when editors decide which scientists to blame. The *Bullied Into Bad By Science* Campaign was started in 2017 by "postdocs and a reader in the humanities and sciences at the University of Cambridge" who were **concerned about the desperate need for publishing reform** to increase transparency, reproducibility, timeliness, and academic rigour of the production and dissemination of scholarly outputs [31] Because, they say (and cite evidence) that early career researchers (ECRs) "**are often pressured into publishing against their ethics** through threats that we would not get a job/grant unless we publish in particular journals". Their petition has a amendment added by Anne Schell (not necessarily endorsed by the initial signers) that specifically addresses integrity: "Stop pressuring ECRs into conducting/writing up underpowered, non-preregistered, p-hacked, HARKed studies. In other words: stop teaching/advising/pressuring people to mutilate data into a 'publishable' form when that distances it from actual science." Preregistration can indeed prevent a lot of this pressure. Pressure can come **after** preregistration, where data could manipulated to conform with the preregistered hypothesis. This highlights the need to hash data **as it is created**. Researchers can unilaterally post (OSF, github, or a library) hashes of their data without jeopardizing the privacy of the scientific group. Later, if asked to manipulate data, this scientist could point to the 3rd party hashes and explain how easily they would be caught: legal and cryptographic checkmate.

In the summer of 2019 at the University of Florida, Computer Architecture PhD candidate Huixiang Chen committed suicide–according to the suicide note posted by his friends and academic colleagues–because he saw no way out of a dispute over the integrity of data [28]. Despite the extra attention this action has elicited, this is not a good way to resolve an academic dispute; this is also precisely the type of situation an integrity system should try to prevent. In no way does the mention of this tragedy imply the guilt or innocence of the people involved. Chen claimed that a joint paper that had been accepted included experiments that had never been conducted, so he was tasked "to make up for the missing experiments" [28]. If journals, funders, departments,

or universities required that experiments be hashed and cashed with a 3rd party when they are created; and the peer reviewers checked the *Pli Cacheté* before acceptance for publication; then disputes of this nature would be resolved easily or never happen at all. As the International Journal of Medical Journal Editors states, "Perceptions of conflict of interest are as important as actual conflicts of interest," [29] similarly, perceptions of misconduct are almost as important as actual misconduct. Openness and authenticated documentation protects everyone from accusations and perceptions of misconduct.

In North America, the Scientific Integrity Consortium aims to form a comprehensive approach to integrity in science, focusing on the best practices for the bureaucracies that manage scientists. A *Pli Cacheté* can complement their work because they identify an "urgent need to refocus the scientific communitys efforts on policing itself" [14]. Anyone can choose to use a unilateral *Pli Cacheté*, a form of policing oneself that prevents investigations altogether. While requiring it would be a bureaucratic solution, it minimizes the need for bureaucratic monitoring and control.

9. Cryptographic Techniques

Consider the backwards problem to evaluate integrity solely on the basis of the data and hypotheses revealed just when an article is published. Potential integrity problems include: 1) computational bugs or errors introduced during data storage, processing or analysis; 2) intentional changing of the data points to affect the final result; and 3) data plagiarism (possibly with data owned by the scientist but dishonestly reused).

9.1. Secure Sketches and FIBE

To easily defeat a hash check for plagiarism, change any one thing, then the checksums will be different. Shams-e-Qays might have called this type of plagiarism "flayed" data. One solution is to borrow a method from biometric passwords called Fuzzy Identity Based Encryption (FIBE). Typical passwords and hashes must be exact to gain access. But a fingerprint scan, for example, is a huge data file with an error deviating from the "true" fingerprint[30]. Instead, one FIBE method creates a **secure sketch** for the older data set with a chosen tolerance for deviation. If both both data sets can unlock the secure sketch, then they are indeed very close and candidates for investigation[30].

These techniques "apply not just to biometric information, but to any keying material that, unlike traditional cryptographic keys, is (1) not reproducible precisely and (2) not distributed uniformly." Thus FIBE can also be used to honestly evaluate data that contain slight variations when reproduced. The **secure sketch** is a ciphertext which "produces public information about its input w that does not reveal w, and yet allows exact recovery of w given another value that is close to w." A **fuzzy extractor** is a similar (near) uniform length ciphertext. Fuzzy cryptographic methods incorporate metric spaces and distance functions where the distance between w and w' is a parameter chosen at encryption time[30]. An enticing idea is a fuzzy data auditing system (prototyped in Li et al [31]) for use in a large multi-grid system [32] but with tolerance for error because "apparently routine data manipulation workflows become rife with mundane complexities as researchers struggle to assemble large, complex datasets." [33].

9.2. Benford's Law

Benford's Law is an empirical observation (mathematical proofs of which are hotly debated) where 1 is the most common first digit, 2 is the 2nd most common, descending monotonically to 9. An unadulterated data set will often demonstrate an exponentially descending Benford Curve. Manual adjustments to data can substantially change the distribution of numeral frequency, serving as an indicator for potential fraud. Two data sets which have exactly the same deviation from the Benford Curve would be another indicator. Benford's Law is particularly useful for detection fraud in data sets for regression[34]. Storing copies of Benford Curves in a *Pli Cacheté* would serve as an additional integrity check, one that–in the long term–would not depend on having a reproducible hash function.

In the detection of plagiarism, Benford's Law and secure sketches have a powerful synergy. While merely one change defeats a plagiarism audit with a checksum, a secure sketch would require many more changes. But with more manufactured changes, the Benford Curve would likely be more erratic–and therefore possibly fraudulent. Defeat a Secure Sketch, get caught by a Benford Curve.

9.3. Resampling Detection

A method to detect stretching and rescaling is the Expectation-Maximization algorithm which "is applied to estimate the interpolation kernel parameters, and a probability map (called *p*-map) that is achieved for each pixel provides its probability to be correlated to neighbouring pixels. The presence of interpolated pixels results in the periodicity of the map, clearly visible in the frequency domain" [35]. This is an argument for the caching of raw data. Bowman and Keene advocate for the preservation of raw data in general because it "will allow the researchers to view the entire spectrum of what was done rather than simply what was reported" [36].

9.4. Auditing Amalgamated Research

A four step plagiarism audit process emerges: 1) Does the checksum match another dataset?; 2)Does the secure sketch match another dataset?; 3) Is the Benford curve erratic?; 4) Is there evidence of resampling if the data is raw? This is a check of a proposed dataset against previously adopted datasets. Passing this, a new hash of the proposed dataset should be checked against the original hash stored with in a *Pli Cacheté* (hopefully from the date of its creation). Failing this check leads to an investigation: 1) If the secure sketch still matches, then the changes are minor; and 2) Are the Benford Curves consistent with a random change, erratic change, or remain a consistent Benford Curve?

10. Conclusions

While a *Pli Cacheté* can be immediately and unilaterally, some promising directions for future research include:

1. Establish an effective and universal secure sketch or fuzzy extractor (like SHA);
2. Find or create a resampling test suitable for scientific data audits;

3. Create an open repository of fraudulent and corrupted data for testing;
4. Audit an already constructed amalgamated research database; and
5. Create a standardized ontology of reproducible science.

The biggest barrier to implementation is the social challenge of convincing a scientific community to adopt any mandatory methods. Herein lies the importance of the backwards problem: as investigations reveal fraud or unintentional errors, then the case for the forward problem of preventing the loss of integrity becomes more compelling.

References

[1] The European Code of Conduct for Research Integrity. 2017. http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf

[2] Sadeghi, R. 2019. The attitude of scholars has not changed towards plagiarism since the medieval period: Definition of plagiarism according to Shams-e-Qays, thirteenth-century Persian literary scientist. Research Ethics. Vol. 15(2) 13. https://doi.org/10.1177/1747016116654065

[3] Laine, C and The Council of Science Editors. 2012. Description of Research Misconduct. Approved March 30th, 2012. https://www.councilscienceeditors.org/resource-library/editorial-policies/white-paper-on-publication-ethics/3-1-description-of-research-misconduct/#312

[4] Coudert, F. 2019. Correcting the Scientific Record: Retraction Practices in Chemistry and Materials Science. Chem. Mater.201931103593-3598. Publication Date:May 28, 2019. https://doi.org/10.1021/acs.chemmater.9b00897

[5] Chawla, D. S. 2019. Retracted chemistry studies most often plagued with plagiarism. Chemical Engineering News. Published: May 31, 2019. https://cen.acs.org/research-integrity/misconduct/Retracted-chemistry-studies-often-plagued/97/web/2019/05

[6] Sean Thorpe, Indrajit Ray, Tyrone Grandison, Abbie Barbir. Formal Hash Compression Provenance Techniques For The Preservation Of The Virtual Machine Log Auditor Environment. The International Journal of Information Science and Computer Application (IJISCA), Vol 1, pp 1-10. https://www.tyronegrandison.org/uploads/1/8/8/1/18817082/newfinalijisca_papercamerarready2012.pdf

[7] Blazic, AJ. Trusted Archive Authority–Long Term Trusted Archive Service. EurOpen Conference Processings, Czech Open Systems Users' Group, pp.107-119 May, 2001. https://www.europen.cz/Anot/30/hlavni.pdf#page=107

[8] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M.B. Jones, K. Kowalik, et al. 2019. Computing environments for reproducibility: capturing the "Whole Tale" Future Gener. Comput. Syst., 94, pp. 854-867 https://doi.org/10.1016/J.FUTURE.2017.12.029

[9] The Swiss Centre of Expertise in the Social Sciences (FORS). Implementation of the CoreTrustSeal for the Repository DARIS. Mar. 20, 2018. https://www.coretrustseal.org/wp-content/uploads/2018/03/DARIS.pdf

[10] Assante, M., Candela, L., Castelli, D. and Tani, A. 2016. Are Scientific Data Repositories Coping with Research Data Publishing? Data Science Journal, 15, p.6. http://doi.org/10.5334/dsj-2016-006

[11] University of Minnesota Libraries. Data Management Plans (DMPs). Retrieved September 8th, 2019. https://www.lib.umn.edu/datamanagement/DMP

[12] Sivagnanam, S., Nandigam, V. and Lin, K. 2019. Introducing the Open Science Chain: Protecting Integrity and Provenance of Research Data. In Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (p. 18). ACM. https://dl.acm.org/citation.cfm?id=3332203 and https://www.opensciencechain.org

[13] Erren TC. 2008. On establishing priority of ideas: revisiting the pli cacheté (deposition of a sealed envelope). Medical Hypotheses 2008;71(1):8e10 https://doi.org/10.1016/j.mehy.2008.08.013

[14] Kretser, A., Murphy, D., Bertuzzi, S. et al. 2019. Sci Eng Ethics. 25: 327. https://doi.org/10.1007/s11948-019-00094-3

[15] Editorial Board. Research integrity is much more than misconduct. 2019. Nature 570, 5. https://www.doi.org/10.1038/d41586-019-01727-0

[16] United States Patent and Trade Office (USPTO). Provisional Application for Patent. Retrieved September 8th, 2019. https://www.uspto.gov/patents-getting-started/patent-basics/types-patent-applications/provisional-application-patent

[17] World Intellectual Property Organization (WIPO). Retrieved September 8th 2019. Summary of the Paris Convention for the Protection of Industrial Property (1883). https://www.wipo.int/treaties/en/ip/paris/summary_paris.html

[18] Tracy, J. E. 1939. The introduction of documentary evidence. Iowa L. Rev., 24(3), 436-463. https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/ilr24&id=458&men_tab=srchresults

[19] Graham, M.A. 1994. Cheating at Small Colleges. Journal of College Student Development, 35(4), pp.25560. https://eric.ed.gov/?id=EJ489082

[20] York University, Canada. Retreived September 8th, 2019. Academic Dishonesty in Laboratory Environments. https://teachingcommons.yorku.ca/resources/teaching-strategies/academic-integrity/academic-dishonesty-in-laboratory-environments/

[21] Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., Frank, M. C. 2018. A Practical Guide for Transparency in Psychological Science. Collabra: Psychology, 4(1), 20. http://doi.org/10.1525/collabra.158

[22] Center for Open Science. 2018. Guidelines for Transparency and Openness Promotion in Journal Policies and Practices "The TOP Guidelines." Version 1.0.1 https://osf.io/9f6gx/

[23] Mohar, David and Glasziou, Paul. 2019. How can we improve organizational assessment of researchers? World Conference on Research Integrity, Hong Kong, China. http://www.wcri2019.org/uploads/files/archive_plenary/day_4_june_5/concluding_plenary_david_moher_and_paul_glasziou_hong_kong_principles_final_plenary_session.pdf

[24] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. 2005. A survey of data provenance in e-science. SIGMOD Rec. 34, 3 (September 2005), 31-36. https://doi.org/10.1145/1084805.1084812

[25] Scargle, J.D., 2000. Publication Bias: The File-Drawer Problem in Scientific Inference. Journal of scientific exploration, 14(1), p.91. https://arxiv.org/abs/physics/9909033

[26] Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. 2008. Selective Publication of Antidepressant Trials and its Influence on Apparent Efficacy. N Engl J Med; 358:252-260. https://doi.org/10.1056/NEJMsa065779

[27] Logan, C., et al. 2017. Retrieved September 8th, 2019. http://bulliedintobadscience.org

[28] Anonymous. 2019. The Hidden Story Behind the Suicide PhD Candidate Huixiang Chen. Medium, June 29th, 2019. https://medium.com/@huixiangvoice/the-hidden-story-behind-the-suicide-phd-candidate-huixiang-chen-236cd39f79d3

[29] International Committee of Medical Journal Editors. Website Viewed September 8th, 2019. Conflicts of Interest. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/author-responsibilities--conflicts-of-interest.html

[30] Y Dodis, R Ostrovsky, L. Reyzin, A Smith. 2008. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. A preliminary version appeared in Eurocrypt 2004 [DRS04]. SIAM Journal on Computing, 38(1):97139 http://web.cs.ucla.edu/~rafail/PUBLIC/89.pdf

[31] Y. Li, Y. Yu, G. Min, W. Susilo, J. Ni and K. R. Choo, 2019. Fuzzy Identity-Based Data Integrity Auditing for Reliable Cloud Storage Systems. IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 1, pp. 72-83, 1 Jan.-Feb. https://doi.org/10.1109/TDSC.2017.2662216

[32] G. Tylissanakis and Y. Cotronis, "Data Provenance and Reproducibility in Grid Based Scientific Workflows," 2009 Workshops at the Grid and Pervasive Computing Conference, Geneva, 2009, pp. 42-49. https://doi.org/10.1109/GPC.2009.16

[33] K. Chard et al. 2016. I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. 2016 IEEE International Conference on Big Data, Washington, DC, 2016, pp. 319-328. https://doi.org/10.1109/BigData.2016.7840618

[34] Jamain, A. Benford's Law [thesis]. 2001. Imperial College of London and Ecole Nationale Superieure. https://wwwf.imperial.ac.uk/~nadams/classificationgroup/Benfords-Law.pdf

[35] Piva, A. An Overview on Image Forensics. ISRN Signal Processing, vol. 2013, Article ID 496701, 22 pages, 2013. https://doi.org/10.1155/2013/496701

[36] Bowman, N. D. and Keene, J. R. 2018. A Layered Framework for Considering Open Science Practices. Communication Research Reports, 35:4, 363-372. https://doi.org/10.1080/08824096.2018.1513273