Privacy is an integral part of our society. However, openness and availability of information is equally important for scientific advancements. The fine line between ethical responsibility and evolving re-computational practices should be clearly defined. Re-Computation (ER) is an initiative by [CreditsHere] academic community to enable early stage researchers benefit from existing experiments without reinventing the wheel. Academic Re-Computation (ARC) will allow researchers to advance their investigation while integrating and utilizing the existing solution with research efficiency improvements and cost effective experimentations.

ARC emphasis on set of practices for any given experiment which include software (application programming) as a source of conclusion to the discussed problem. This includes software, dataset, parameters, pre and post conditional legends and documentation to re-produce the identical results while executing the same experiment. Published work should serve complete and ultimate documentation of the experiments. In situation where format of publishing journal imposes restrictions, format (space/style) or the detailed information on experiment would cause distraction from main presented idea, adequate information should be provided in terms of index, reference and persistent web links.

The provided solution should be compete with all its dependencies including documentation, libraries, data files, database schema, licencing information and source code. The whole process to rerun the experiment should be explained in detail and step by step. Various techniques, during discussion at summer school, to share the source code and data were discussed. One solution is to utilize virtualization technique to prepare a virtual machine with operating system, development tools and actual experiment. The benefit of this approach is ease to re-compute the experiment with minimum setup overhead. Setup process can be time consuming and costly, finding the right version of required software itself can be challenging job. Disadvantage of virtual machine involves licensing issues of software. Another approach is to provide all the resources in terms of documentation, installation procedures and details to setup the experiment environment.

The details to re-produce the experiment should protect sensitive information and identities in all forms. This also includes business or organizational secrets and intellectual proprietary licences. David Leoni has presented an interesting example with minimal details of personal information leading to person’s identification. Study has proved that 87% of American citizens can be uniquely identified just by knowing their gender, zip code and birth place [1]. Sufficient study has been done and still evolving to anonymities the data before publishing in public domains. We intend to provide awareness (ethics, privacy and potential risks) among the researchers who wish to publish their work under the umbrella of Academic Re-computation initiative.

Normally, researcher themselves do not participate in unethical and vulnerable practices. However, their provided data and implementation can cause harm in hands of adversaries if strict measurements are not in place. According to 2014 Information Security Survey conducted by PricewaterhouseCoopers the overall cost of security breaches for large and small scale business is increasing. The number of security breaches for large organization is 81% whereas 60% of small businesses were affected in same time [2]. Although there is a slight decline in overall security incidents but numbers are still alarming. We advocate a common code of conduct including implementation and data sharing.

Ethical and privacy concerns are the way points how ARC will be implemented, we summarize and discuss these issues in following three categories.

- Disclosure or Identity Theft: Discloser is a privacy threat which produces back-track to original experimental source. The identity of users should be concealed while providing pseudo-users instead of real users. Anonymity of data and experiment procedures gain more significance when there is human intervention part of experiment in direct or remote sense. Published ARC should follow the predefined user rights [User-Acceptance-agreements]. Proven data mask techniques should be used to ensure that no adversary would be able to track the users. Data mask is a technique which allows changing the data without changing its original format. The used mechanism to mask the data should include complete set of data and as well individual and independent dataset.

- Integrity: The integrity threat involves intentional or unintentional change in data leading to sensitive information. The details provided in ARC should not lead to identification and sufficient mechanism to obscure the identities must be fundamental practise before sharing the user information. Provided data should never include sensitive and personal details in any forms.

- Replication: ARC should provide means to reproduce the results while cloning the data and parameters involved in experiment. But, it is important that replication should not imitate to possible user scenarios. For example, names of users from social data replaced with random number generator but other delicate details left intact. Researchers should ensure that replication of data should not lead to real identity or Intellectual property infringement.

References:

[1] David Leoni. 2012. Non-interactive differential privacy: a survey. In Proceedings of the First International Workshop on Open Data (WOD '12). ACM, New York, NY, USA, 40-52. DOI=10.1145/2422604.2422611 http://doi.acm.org/10.1145/2422604.2422611

[2] http://www.pwc.co.uk/assets/pdf/cyber-security-2014-exec-summary.pdf