How can photo sharing inspire sharing genomes?

Vinicius V. Cogo[1], Alysson Bessani[1], Francisco M. Couto[1], Margarida Gama-Carvalho[2], Maria Fernandes[3], and Paulo Esteves-Verissimo[3]

[1] LaSIGE, Faculdade de Ciências, Universidade de Lisboa, PT
[2] BioISI, Faculdade de Ciências, Universidade de Lisboa, PT
[3] SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, LU

Abstract. People usually are aware of the privacy risks of publishing photos online, but these risks are less evident when sharing human genomes. Modern photos and sequenced genomes are both digital representations of real lives. They contain private information that may compromise people's privacy, and still, their highest value is most of times achieved only when sharing them with others. In this work, we present an analogy between the privacy aspects of sharing photos and sharing genomes, which clarifies the privacy risks in the latter to the general public. Additionally, we illustrate an alternative informed model to share genomic data according to the privacy-sensitivity level of each portion. This article is a call to arms for a collaborative work between geneticists and security experts to build more effective methods to systematically protect privacy, whilst promoting the accessibility and sharing of genomes.

Keywords: privacy, data sharing, biology and genetics

1 Introduction

We live in a world plenty of connected devices and services that stimulate and simplify data sharing, which promote the acceptance of exposure risks. Nowadays, the general public recognizes several privacy risks in sharing photos on the Internet. This was promoted by the public widespread dissemination of some information leakages that caused severe privacy harms, which made users start to demand more privacy guarantees to continue sharing their data on online platforms [14].

Solutions for photo sharing already faced several privacy-related conflicts and policies changes, and life sciences can learn from them. An analogy between sharing photos and genomes may increase people's awareness on privacy risks and contributes to avoid future leakages that could damage people's willingness to share genomic data. We emphasize that this comparison is reasonable since sequenced genomes and modern photos are digitized records of real lives. Both contain private information that may compromise people's privacy, and most of the times, their highest value is only achieved when shared with others.

Human genome is privacy sensitive since it contains personal information, and researchers need the access to large collections of genomes to accelerate medical breakthroughs. The ethical appeal for disclosure stimulates altruistic individuals to donate biological samples for medical and genomic research. However, this point of view must coexist with the ethical discussion on the risks to donors' privacy and encourage the development of secure models to share genomic data [1].

Privacy and data sharing are not mutually exclusive. Properly discussing and defending privacy encourages the responsible data sharing and extends donors' engagement and trust in researches. Recent publications corroborate with the ideas that clearly informing donors about the privacy risks of their choices does not affect negatively their willingness in donating samples [12], and that there is a need for balancing data access and privacy in genomics [19].

In this article, we propose an analogy between privacy aspects of sharing photos and sharing genomes, which contributes to clarify the privacy risks in the latter. Additionally, we illustrate possible advances in sharing genomes with an alternative informed model to share genomic data according to the privacy-sensitivity of their portions. These two contributions promote the accessibility and sharing of human genomes, whilst advocates their responsible management considering the privacy of sample donors.

2 An analogy between sharing photos and sharing genomes

We defined an analogy by comparing the similarities and features of the processes of sharing photos and sharing genomes, which is based on the following aspects:

- Some portions of data are more privacy-sensitive than others.
- One's data may affect the privacy of others.
- Systematically detecting the privacy-sensitive portions of data is feasible.
- After classifying the portions, decide how to share them.
- The impact of data sharing is unpredictable.

On each topic, we first describe it from the perspective of sharing photos and then we present the analogy on how does it apply in sharing human genomes. Note the present analogy is non-exhaustive since further discussions from the community may identify other similarities in the future.

2.1 Some portions of data are more privacy-sensitive than others

Some elements in photos (e.g., faces and places) may disclose sensitive information about the people that own or are depicted in them, such as identity, ancestry, health, behavior, preference, possession, and location. Similarly, genomes contain portions of sequences that contain more critical information (e.g., predisposition to a disease, parental correlation) about their donors and their relatives [15].

Authors of recent publications managed to compromise donors' privacy by targeting specific portions of human genomes, such as short-tandem repeats [7], disease-related genes [16], and genomic variations [8]. These elements may disclose information, for example, related to identity, ancestry, and health.

2.2 One's data may affect the privacy of others

Photos portraying other individuals may compromise their privacy, as well as photos containing elements related to controversial topics may affect the privacy and safety of owners' relatives (e.g., [10]). In human genomes, some information is hereditary (e.g., Y chromosome from father to son), and thus compromising the privacy of one subject genome can also affect his relatives [15].

2.3 Systematically detecting the privacy-sensitive portions of data is feasible

Detecting the privacy-sensitive elements in photos includes recognising faces [9], activities [17], texts [11], signs and other location-specific elements [6]. Recently, we proposed a method that detects the privacy-sensitive portions of human genomes by comparing small DNA portions against a knowledge database of privacy-sensitive genomic sequences [5]. In both cases (photos and genomes), the detection compares small elements against large databases of known patterns. Although those detection methods contribute to privacy protection by differentiating sensitive information, the challenges remain mostly in building comprehensive knowledge databases and querying them efficiently.

2.4 After classifying the portions, decide how to share them

Regarding photo sharing there are two distinct options: (1) enable the share if the person concludes it does not compromise his/her privacy nor the privacy of others, and (2) share a portion of the photo, which the person believes it does not compromise anyone's privacy, while keeping private or obfuscating the remaining sensitive portions for the general public. Excluding these two options there is always the possibility to not share the photo. Recent publications proposed alternative informed models to share photos considering the privacy-sensitivity of their portions [9, 18]. Similar to photos, every human genome contains some privacy-sensitive portions. We advocate that sharing certain portions of data is more attractive than sharing nothing, and those privacy-sensitive portions may still be shared in a controlled way (e.g., using the cryptographic methods discussed in [15]). In the next section, we propose an alternative informed model to share genomic data considering the privacy-sensitivity of their portions.

2.5 The impact of data sharing is unpredictable

Sharing photos may have an immediate impact in the lives of a small number of people related or depicted on them. However, the global impact of a shared photo is unpredictable. For instance, a photo can be considered meaningful to history independently from depicting everyday-life or epic moments. Additionally, several quotidian applications we use rely on common user-contributed content, as well as some news we read depend on participatory journalism. The contribution of sharing each data is little, but all these incremental collaborations have a huge impact. The same happens with human genomic data, where the highest value of photos and genomes is most of times achieved only when sharing them with others. The individual altruism in contributing to medical and genomic studies has an extreme importance on the breakthroughs in health-related areas.

3 An informed model to share genomic data

With all the previously mentioned aspects in mind, we call attention to the opportunities a hybrid solution can bring to balance data access and privacy of genomic data [5, 19]. Our proposal is to use the referred detection method [5], as mentioned in §2.3, to identify and differentiate the privacy-sensitive sequences of human genomes from the remaining portions. This enables one to keep the small privacy-sensitive portions (i.e., less than 12%, conservatively [5]) of human genomes under a strict access control list, and make the remaining portions directly accessible to researchers and projects, according to the rights defined at their registry in the data repository. The completeness of this method evidences that there is already a large body of knowledge on the privacy sensitiveness of human genomes and that the discovery of novel privacy-sensitive sequences is unlikely using current methods (e.g., [7, 8, 16]). In this section, we introduce this model and its main internal components, as well as place it in the ecosystem and describe our vision on how should players interact with it.

3.1 Players and Interactions with the Model

There are four main players in the ecosystem of genomic data sharing. *Sample donors* donate biological material to a sample manager and inform their preferences on data sharing (if any). *Sample managers* receive, manipulate, sequence, store, and provide these biological specimens and their resulting data. Research projects are study proposals, encompasses one or more researchers, and have well-defined goals that require access to data associated with specific samples. *Researchers* are entities within projects that consume data from the storage system according to donors preferences and other permission rules. *Auditors* are stakeholders (e.g., governments, investors, donors, and data managers) that want to verify when and which researchers accessed specific data sets.

Donors fill consent forms at their registry to comply with regulations and to inform their preferences on data sharing. Donors should be free to customize their informed preferences to state they want to automatically participate in projects related to specific topics (i.e., a blanket consent). They should also inform they want to contribute with their samples to additional specific projects they sympathize with (i.e., opt-in). Additionally, donors could delegate the decision of participating in which projects to data controllers acting on behalf of groups of individuals. Exceptionally, donors could separately forbid the use of the non-sensitive portions of their genomes by specific projects they disagree, or may require to re-categorise some non-sensitive portions as privacy-sensitive (i.e., opt-out). The per-project opt-out dissuades an eventual retraction of all genomes from the platform if an isolated misuse happens [4]. When a donor dies, the sharing preferences may become open or be kept the same, while his/her relatives gain the ability to explicitly customize them.

In the envisioned model, researchers should register themselves in the system and propose projects that are approved in the same way and with the same responsibilities it is currently done in biobanks and other repositories. Projects (i.e., groups of researchers) may start working with all non-sensitive sequences immediately, must wait for a short period to start using the automatically authorised privacy-sensitive portions, and have the option to request access to the privacy-sensitive portions of other genomes of interest. The utility of sequenced data is kept intact to authorised researchers in this model, which complements other approaches from the literature (e.g., [2]).

3.2 Internal Components

This data sharing model can be adapted to different legal, geographic, and organizational regulations. Additionally, this model, as depicted in Figure 1, is completely independent of the protocols and technologies necessary to implement it. In the following, we describe four components that are of extreme importance to this model, but others can be integrated to them if needed in the future.

Evolution Module. The knowledge database from the DNA privacy detector can be automatically updated to address future attacks as new privacy-sensitive sequences are identified [5]. Thus, the detection method is generic and evolvable—i.e., it does not become outdated since public databases can be automatically tracked for updates as they evolve. An *evolution module* in this system architecture should allow the stored data sets to be re-analyzed at any moment and attested again for their privacy-sensitivity. As soon as a new privacy-sensitive sequence is identified, the data sets updated, access rules are adapted accordingly, and the access history is logged for future inquiries.

Storage. *Storage* components should retain and provide the large amount of genomic data coming from life-sciences institutions. Storage infrastructures encompass several options from private data centers to public clouds. Data from human genomes in the envisioned model is stored according to the privacy-sensitivity of its portions. The privacy-sensitive portions of human genomes must be stored in infrastructures with appropriate levels of security and dependability, while the non-sensitive ones can stored in more affordable infrastructures. Noticeable, this hybrid model improves the cost efficiency of any storage system since it reduces the percentage of data requiring strong security and dependability premises.

The level of security and dependability depend on the use of encryption, information dispersal, data replication, etc. Choosing the best fit is orthogonal to this model and depends also on the legal constraints defined by regulators

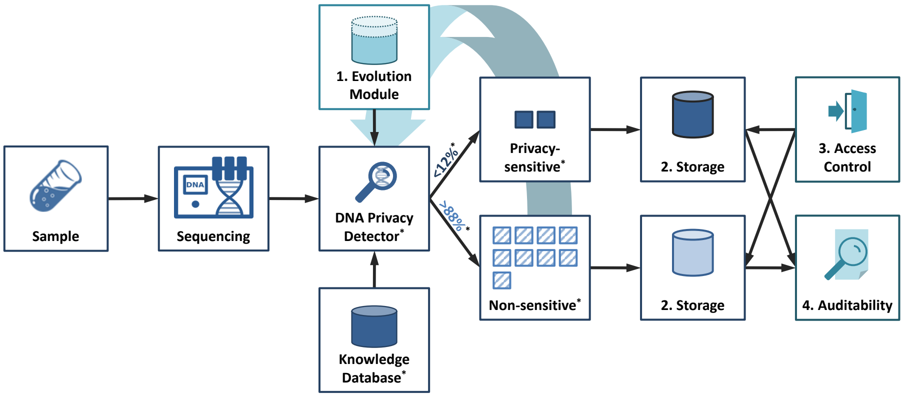

Fig. 1. Overview of the hybrid data sharing model. This model considers that genomes have their privacy-sensitive portions differentiated* from the remaining ones.

from the region of the sample manager. Restrictive regulations may impede sending data to infrastructures in other countries, while less restrictive ones may allow the use of standard encryption and public clouds. For instance, the storage solution from the BiobankCloud project already considers this range of options and provides data storage in private repositories, in single public clouds, and in multiple clouds (i.e., a cloud-of-clouds) [3].

Access Control. Access control establishes a differential access to users, accordingly to their roles and analysis. An *access control* solution should verify and permit researchers to access the different portions of the genomes they are allowed to. Additionally, the access control complements the evolution module by automatically updating the lists and rules according to the data sets' version.

There are three main factors to authenticate an access request: something the user knows (e.g., a password), has (e.g., a token), or is (e.g., biometrics). Combinations of them can be used to increase the difficulty for an illegitimate user having access to a resource. For instance, the BiobankCloud platform [3] requires each user to authenticate with his/her password and an one-time password generated using a mobile phone or a Yubikey. Additionally, cryptographic solutions complement access control mechanisms since an attacker that circumvents the access control does not obtain the data in clear.

Auditability. Auditability is the relative ease of auditing a system or an environment, acts as a deterrent measure, and complements preventive ones, such as security, dependability, and privacy-protection. An *auditability* component should enable stakeholders to assess at any moment exactly who accessed what data in a chronological order. Auditors should access only some metadata about the files, the access logs, and access control rules—i.e., they do not need to access the whole data sets of genomic data. The auditability component complements the evolution module by allowing the detection of who has read previous versions of a data set that was re-analyzed because it could contain previously unknown privacy-sensitive sequences. Accountability supplements auditability by ensuring all actors and actions performed on the data have been persistently recorded as evidence [13]. The system must keep an indelible tamper-proof track of data accessed by researchers, in order to detect, analyze, and sanction misuses.

4 Final Remarks

In this work, we presented an analogy between privacy aspects of sharing photos and sharing genomes, and proposed an informed model to share genomic data according to the privacy-sensitivity of their portions. The analogy contributes to advancing the privacy-perception in sharing genomes by comparing it to some well-known examples and threats from sharing photos. The informed model motivates the discussion of novel solutions for sharing genomic data considering their privacy-sensitivity.

Notwithstanding, there are many open questions (related to this model and the problems identified in the analogy) that deserve further investigation and discussion within the community, namely:

- How to provide this data sharing model without incurring in unreasonable increased management effort?
- How can public clouds be securely used in this model to reduce the costs of creating and maintaining private storage infrastructures (e.g., in biobanks)?
- Which additional type of genomic data, beyond those discussed in [5], can be considered privacy-sensitive and should thus be detected?
- There are methods that associate genomic information and photographic records (e.g., selecting individuals in a database using the association between specific SNPs and the probability of an individual having brown or blue eyes [20]). Understanding the impact of those associations on subjects' privacy may contribute for more complete protection methods.

Currently, there is a great investment to advance from conventional to precision medicine, which can succeed only if we embrace genomic data sharing in a secure and controlled environment. This article is a call to arms for geneticists and security experts, to work together and build better and more effective methods to systematically protect privacy, whilst improving the accessibility and sharing of genomic data.Our model can even be accommodated in a linked data or beacon service perspective, sharing sensitive data only means that we need to be aware of what and how we share to make it safe and useful for everyone.

Acknowledgements. This work was partially supported by the Fundação para a Ciência e para a Tecnologia, through the LaSIGE (UID/CEC/00408/2013) and BioISI research units (UID/MULTI/04046/2013), by the Fonds National de la Recherche Luxembourg (FNR) through the PEARL grant FNR/P14/8149128, and by the European Commission, through the BiobankCloud (FP7-ICT-317871) and SUPERCLOUD (H2020-ICT-643964) projects.

References

[1] Allen, A.L.: What must we hide: The ethics of privacy and the ethos of disclosure. Thomas L. Rev. 25, 1 (2012)

[2] Ayday, E., Raisaro, J.L., Hengartner, U., et al.: Privacy-preserving processing of raw genomic data. In: Data Privacy Management and Autonomous Spontaneous Security, pp. 133–147. Springer (2014)

[3] Bessani, A., et al.: Biobankcloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. In: Proc. of the DMAH 2015 (2015)

[4] Brenner, S.E.: Be prepared for the big genome leak. Nature 498(7453), 139–139 (2013)

[5] Cogo, V.V., Bessani, A., Couto, F.M., et al.: A high-throughput method to detect privacy-sensitive human genomic data. In: Proc. of the 14th ACM Workshop on Privacy in the Electronic Society. pp. 101–110. ACM (2015)

[6] Doersch, C., Singh, S., Gupta, A., et al.: What makes paris look like paris? ACM Transactions on Graphics 31(4) (2012)

[7] Gymrek, M., McGuire, A.L., Golan, D., et al.: Identifying personal genomes by surname inference. Science 339(6117), 321–324 (2013)

[8] Homer, N., Szelinger, S., Redman, M., et al.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet 4(8), e1000167 (2008)

[9] Ilia, P., Polakis, I., Athanasopoulos, E., et al.: Face/off: preventing privacy leakage from photos in social networks. In: Proc. of the CCS'15. pp. 781–792. ACM (2015)

[10] Jones, S., Norton-Taylor, R.: 'Congrats to Uncle C'—how his wife's Facebook page exposed new MI6 head. The Guardian (Jul 2009)

[11] Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern recognition 37(5), 977–997 (2004)

[12] Kaufman, D.J., Murphy-Bollinger, J., Scott, J., et al.: Public opinion about the importance of privacy in biobank research. The American Journal of Human Genetics 85(5), 643–654 (2009)

[13] Ko, R.K.: Data accountability in cloud systems. In: Security, Privacy and Trust in Cloud Systems, pp. 211–238. Springer (2014)

[14] Kokolakis, S.: Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & Security (2015)

[15] Naveed, M., Ayday, E., Clayton, E.W., et al.: Privacy in the genomic era. ACM Computing Surveys (CSUR) 48(1), 6 (2015)

[16] Nyholt, D.R., Yu, C.E., Visscher, P.M.: On Jim Watson's APoE status: genetic information is hard to hide. Eur. J. Hum. Genet. 17, 147–149 (2009)

[17] Poppe, R.: A survey on vision-based human action recognition. Image and vision computing 28(6), 976–990 (2010)

[18] Ra, M.R., Govindan, R., Ortega, A.: P3: Toward privacy-preserving photo sharing. In: Proc. of the USENIX Symposium on NSDI'13. pp. 515–528 (2013)

[19] Vayena, E., Gasser, U.: Between openness and privacy in genomics. PLoS medicine 13(1), e1001937 (2016)

[20] Walsh, S., Liu, F., Ballantyne, K.N., et al.: Irisplex: A sensitive {DNA} tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Science International: Genetics 5(3), 170–180 (2011)