Ethics-aware data governance (Vision Paper)

Letizia Tanca[1], Paolo Atzeni[3], Davide Azzalini[1], Ilaria Bartolini[2], Luca Cabibbo[3], Luca Calderoni[2], Paolo Ciaccia[2], Valter Crescenzi[3], Juan Carlos De Martin[6], Selina Fenoglietto[6], Donatella Firmani[3], Sergio Greco[4], Francesco Isgrò[5], Dario Maio[2], Davide Martinenghi[1], Maristella Matera[1], Paolo Merialdo[3], Cristian Molinaro[4], Marco Patella[2], Roberto Prevete[5], Elisa Quintarelli[1], Antonio Santangelo[6], Andrea Tagarelli[4], Guglielmo Tamburrini[5], and Riccardo Torlone[3]

[1]Politecnico di Milano, Italy, email: *first.last*@polimi.it
[2]Università di Bologna, Italy, email: *first.last*@unibo.it
[3]Università Roma Tre, Italy, email: *last*@uniroma3.it
[4]Università della Calabria, Italy, email: *last*@dimes.unical.it
[5]Università di Napoli Federico II, Italy, email: *last*@unina.it
[6]Politecnico di Torino, Italy, email: *last*@polito.it

Abstract. The number of datasets available to legal practitioners, policy makers, scientists, and many other categories of citizens is growing at an unprecedented rate. Ethics-aware data processing has become a pressing need, considering that data are often used within critical decision processes (e.g., staff evaluation, college admission, criminal sentencing). The goal of this paper is to propose a vision for the injection of ethical principles (fairness, non-discrimination, transparency, data protection, diversity, and human interpretability of results) into the data analysis lifecycle (source selection, data integration, and knowledge extraction) so as to make them first-class requirements. In our vision, a comprehensive checklist of ethical desiderata for data protection and processing needs to be developed, along with methods and techniques to ensure and verify that these ethically motivated requirements and related legal norms are fulfilled throughout the data selection and exploration processes. Ethical requirements can then be enforced at all the steps of knowledge extraction through a unified data modeling and analysis methodology relying on appropriate conceptual and technical tools.

1 Introduction

Traditional knowledge extraction (search, query, or data analysis) systems hardly pay any specific attention to ethically sensitive aspects and to the ethical and social problems their outcomes could bring about. However, such aspects are now becoming prominent, especially with regard to the protection of fundamental human rights and their underpinnings in normative ethics [14]. These demands are broadly reflected into codes of ethics, in the Responsible Research and Innovation (RRI) approach of the European Commission HORIZON 2020,

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy.

in statements of the European Group on Ethics in Science and Technology (EGE, https://ec.europa.eu/research/ege) - with regard to research and innovation in the area of automated data selection and exploration processes - and also in legally binding regulations such as the EU General Data Protection Regulation (GDPR, https://www.eugdpr.org). More recently, computer scientists themselves - via professional organizations such as ACM and Informatics Europe - have been stressing the importance of raising, within their discipline, awareness with respect to ethical and societal issues regarding the use of data [18, 34]. Coherently with this broad ethical and legal framework, we propose a vision aiming to enhance and verify the protection and advocacy of these rights and values throughout each step of the knowledge extraction chain, thus providing all involved stakeholders with a set of novel computational methodologies and techniques to protect and promote fundamental ethical guarantees in data governance. Realizing this vision needs to be achieved in a principled way by analyzing the ethical challenges that must be addressed, by properly amalgamating and resolving contrasts between various ethical demands (e.g., transparency vs protection), and by developing an extensive list of ethical desiderata for data protection and processing. The resulting Ethical CheckList (ECL) will constitute the high-level specifications for any project embracing this vision.

Accordingly, a preliminary goal is to analyze and clarify the relevant meanings of ethically motivated desiderata about data processing. These notably include transparency, interpretability and understandability, in addition to non-discrimination (fairness), diversity protection and their ethical underpinnings. Moreover, the conceptual relationships between these desiderata and their mutual tensions needs to be analyzed, with the aim of identifying acceptable trade-offs for integrating ethical policies in data selection and exploration processes. This analysis may be used to identify ethically motivated specifications for the various models, technologies and tools to be delivered. With this, our vision will therefore contribute to establish higher social standards for transparency, privacy protection, fairness and non-discrimination in data governance.

The observation and enforcement of ethical principles in data management are achieved by considering the following steps of the analysis lifecycle: *i)* source selection, *ii)* data integration, and *iii)* knowledge extraction. Ensuring that the ECL is applied during all such phases allows all stakeholders to enact law regulations and ethical principles and verify that these are respected. Our vision will hence give birth to an ethical data analysis methodology and associated methodological and technical tools that will enforce the ECL specifications throughout the three above mentioned steps of the analysis lifecycle.

2 Methodology

Our vision's methodology builds on methods and tools for data management and on recent studies on data source selection, data integration, and knowledge extraction. The overall approach is to tackle problems that have a practical significance, providing general methods as well as concrete tools that demonstrate the approach. The relevant activities can be developed on two parallel tracks.

On one track, a new, ethics-aware, knowledge extraction lifecycle is carried on, where the experts in ethical, legal, and social disciplines work side-by-side with the computer scientists to *i)* identify the ethical requirements, *ii)* produce the desiderata checklist and *iii)* identify the most appropriate modeling tool(s) to combine them with the application requirements into a coherent framework. The other track is devoted to the study of novel methods for data source selection, integration, and knowledge extraction that comply with the identified ethical requirements. The two tracks are bridged by a conceptual model, based on the Context Dimension Model (CDM) of [2] described below, able to express the ethical requirements of the ECL by associating the various ethical dimensions (i.e., fairness, transparency, diversity, and data protection), along with their different inflections and levels of enforcement, with the knowledge extraction lifecycle activities. Providing a formal model of ethical requirements, to be adopted in each specific situation of use, will contribute to mitigate or remove possible biases from the considered data management methods. Since at present **no such integrated solutions for dealing with ethical issues in data management exist**, our vision will provide a methodological and technological breakthrough and a concrete answer to issues that the scientific community has started investigating in recent years [30, 25, 10] (`http://wp.sigmod.org/?p=1900`). Existing notable approaches, such as [27], consider scenarios in which there is no control on the knowledge extraction lifecycle, thus dealing with the specific, orthogonal goal of discovering bias in online information derived by third parties through the application of knowledge extraction techniques.

3 Ethical issues in the knowledge extraction lifecycle

We now elaborate on how the various issues arising in the knowledge extraction lifecycle are to be dealt with and how they advance the state of the art.

3.1 Source selection

Within this phase we focus on choosing the data source(s) appropriate for the target objectives in terms of quality and satisfaction of ethical requirements (such as trustability, personal rights protection, and fairness) taking into account the differences between categories and favoring the maximization of source diversity. Under this view, source selection is a novel data management problem, motivated by the recent proliferation of data. The research project more related to our goals is SourceSight [29], which proposes a system to interactively discover valuable sets of sources taking into account reliability and data quality, but without considering ethical issues. Along with institutional datasets, data from location-based services, and other kinds of open data, we also consider online social networks, which are nowadays the preferred communication means for information spread and opinion sharing. We address this problem by assessing the ethical requirements provided by the ECL in this phase, e.g., the bias/fairness/authority of the source population. To this end, we plan to design an iterative process in which the ECL is first assessed for each candidate source so as to select the most informative sources for the domain of interest. Innovative

ways of labeling datasets with ethically- and socially-aware metadata are also needed, with the aim of detecting potential limits or flaws early on (e.g., gender imbalances, ethnic or class misrepresentation or underrepresentation).

3.2 Data integration

Within our vision, suitable tools assist the combination of data from different sources and the extraction of real-world entities. This issue are tackled, for data reconciliation, by leveraging existing record linkage methods and, for schema mapping, by tailoring ethically-aware integration views on the basis of the ethical context. Record linkage seeks to identify which objects refer to the same real-world entity, and is fundamental for data integration. Leveraging humans to compare records based on domain knowledge enables high-accuracy linkage in various domains [16]; however, unrestricted human access to data may not be suitable in the presence of sensitive information. For this reason, our goals include the prevention of sensitive informations leaks, and the collection of all the intermediate data transformation that yielded a given integration result, with the goal of enforcing the protection of user data and providing a transparent access to result generation. We also consider review methods for including the user in the data integration loop, for both data and schema reconciliation, so that high-quality integrated views and fine-grained feedbacks on the result can be provided. Merging conflicting information is critical for recent data integration research [9], especially when dealing with web sources and ethical aspects. Although specific source properties computed in the previous source selection step can help in the detection of fake values, the integration step can further contribute to this problem, empowering human experts with ubiquitous collaborative review tools (thus promoting fact-checking culture).

3.3 Knowledge extraction

We focus on various kinds of knowledge extraction methods: 1. Result personalization, 2. Information diffusion and influence propagation in social networks, 3. Explanation models enabling transparency and interpretability of results, 4. Privacy and security in knowledge extraction systems. Note that the research on topics 1. and 2. is related to the modification of existing techniques for knowledge extraction to *guarantee* that the process and results satisfy the appropriate ethical requirements, while topics 3. and 4. study techniques to *enforce and verify* that the ethical requirements be satisfied by the analysis process.

Result personalization. Personalization can be broadly defined as providing an overall customized, individualized user experience by taking into account the needs, preferences and characteristics of a user or group of users [12]. A data personalization method may re-rank the items in a collection to be shown to a user, focus only on items of interest, or recommend additional options. While personalization delivers relevant content, it also polarizes the perspectives and diminishes serendipity, whereas searching for relevant information on large datasets should provide results that are diverse enough [3] to ensure a fair coverage of the available alternatives. Methods that aim to qualify and quantify personalized experiences and their biasing effects have been overlooked in the literature.

Queries aiming to return only the more interesting results can either provide a ranking of objects (top-k queries) or consider some form of dominance to exclude sub-optimal choices (skyline queries [4]). Thanks to a recent contribution [6], free parameters (e.g., weights), are available to fine-tune the behavior of both kinds of query. For any specific dataset, one can characterize the parameter values that guarantee that the output satisfies the required ethical properties (e.g., preserving the distribution of a protected attribute). Research on such issues for ranking queries has provided preliminary results [36], but no study exists yet for skyline queries. Efficient methods are needed to determine such a set of parameter values, to study its stability wrt. a change in input data [31], and to provide metrics to choose among different parameter configurations. This requires analyzing the overhead incurred by methods providing such ethical guarantees, and studying the trade-off between the efficiency and the ethical level of the query process. Such techniques are also useful to ensure that the items of a query result are diverse enough, providing a balanced view of the result space.

The current user context has been adopted as a criterion for personalization by knowledge filtering. Design methods that support dynamic, context-based filtering of pertinent resources [22] facilitate the development of software that takes context into account. In the traditional software lifecycle, the CDM [2] represents the relevant dimensions of context (e.g., current location, situation of use, role of the user), along with a hierarchy of their possible values; any set of dimension values represents a possible context. Depending on the current context, only relevant data are provided to the user. This kind of personalization is naturally coupled with the ethical dimensions described above, also supporting user preferences and recommendations aware of both context and ethics [28].

Information diffusion and influence propagation in social networks. Online social networks (OSNs) are nowadays the preferred communication means for spreading information and sharing knowledge, such as advertising products/services, promoting ideas, sharing opinions. In this regard, influence maximization is central, i.e., to identify k initial influencers that maximize the spread of influence [33, 15]. An important but often overlooked aspect is that success of an information diffusion process might depend not only on the investment-budget (k), but also on the diversity of the initial influencers, as well as of the targets to be influenced. Members of an OSN present two kinds of diversity: *static*, which includes diversity of kind, socio-cultural aspects and other characteristics exogenous to the OSN; *dynamic*, which includes the knowledge, community experience, and shared information acquired over time. The various types of user diversity should be leveraged to push forward research on information diffusion and influence propagation along two main directions: *i)* diversity concerning the targets to be influenced and *ii)* diversity concerning the initial influencers. The former allows us to capture non-discrimination or fairness aspects in the outcome of the diffusion process; the latter enables modeling different triggering stimuli, which intuitively capture utilitarian aspects [10] in terms of marketing principles (e.g., diversification of users skills implies higher productivity). Addressing both fairness and utility opens to opportunities of ethics-preserving information diffusion,

and can support the development of advanced methods around novel perspectives having ethical implications in OSN data analysis. One such perspective is related to the ever increasing phenomenon of fake-news/misinformation spread on the Web: bringing fairness and utility-oriented diversity aspects into fact-checking and misinformation debunking prompts us to develop sophisticated models to handle competitive influence propagation scenarios. There has been little work in diversity in information diffusion and influence propagation [1, 32, 13, 5]. Most of the existing notions of diversity have been developed around structural features of the network, or are based on user profile attributes, but no existing approach proposes diversity-aware solutions in influence propagation.

Explanation models enabling transparency and interpretability of results. Explanation models for results provided by big data transformations and machine learning (ML) systems are needed. The transparency requirement must be balanced with privacy preservation by identifying context-sensitive trade-offs. With the aim of supporting transparency in big data transformations, we consider techniques for data lineage, so as to trace the relationships between input and output data and to identify underlying processes. Data lineage (aka provenance) concerns data origin and transformation history, and enables transparency by supporting explanations of results and processes. However, it may also disclose private or confidential data, and the use of proprietary transformations [11]. Extending provenance techniques to Big Data poses new challenges and opportunities [35]. Providing explanations for ML systems that are often opaque to human beings is a challenge addressed in the emerging XAI (eXplainable AI) research area [20, 23]. In ML classifications, explanation requests are expressed as why-questions: Why were input data associated with class X? [17]. Answers may come in the form of I/O explanations (exhibiting prototypes of the output class) and inner explanations (additionally exhibiting salient components of both input and intermediate processing data). In I/O explanations one often exhibits a single prototype. In real-life, however, a single prototype may not be representative of the entire class, and discarded classification possibilities are important to interpret outcomes. Accordingly, state-of-art tools need to be extended by extracting multiple prototypes for both classification results and discarded classes for Deep Learning networks and other ML systems, e.g., via statistical methods like activation maximization [23] and sparse coding approaches [21]. Similarly, inner explanation tools need to be extended by identifying components of input and intermediate processing data contributing most to classification outcomes.

Privacy and security in knowledge extraction systems. Focusing on methods for data protection oriented to the preservation of privacy when releasing analysis results, we plan to study techniques for trading-off privacy budget for result accuracy. Aspects concerning privacy and security of personal data are increasingly relevant, as also testified by the GDPR, which unifies data protection laws across all European Union members. Several techniques have been developed so far for privacy protection. Differential privacy promises to enable general data analytics while protecting individual privacy, yet existing mechanisms do not support the wide variety of features and sources used in big-data analytics systems [19].

Data anonymization attempts to provide privacy while allowing general-purpose analysis, but recent de-anonymization results [24] proved that it cannot be relied upon. A further technique is homomorphic encryption, which makes it possible to perform certain operations on a ciphertext without decrypting it [26]. Homomorphic encryption techniques were recently coupled with a novel probabilistic data structure, the Spatial Bloom Filter (SBF) [7], designed to secure out location data. This data structure is suitable for any kind of set-based problem [8], and could be conveniently used in a number of applications. We focus on possible applications of the SBF for data protection. For instance, since an individual's interest may be seen as his own membership to a specific set, a promising application of the SBF is related to those services that rely on outsourced data reflecting people's interests for marketing actions and other commercial purposes. We focus on the study of methods for data protection oriented to the preservation of the privacy of individuals when releasing statistics. Indeed, aggregated data can be combined for inferring personal information, even if randomized.

4 Outlook

Embracing our vision will contribute to establish higher standards in democratic societies for transparency, privacy protection, fairness and non-discrimination in data governance. Trust building and public confidence will be fostered by supporting scrutiny without violating privacy, and by reducing the opaqueness of current data technologies. The resulting competencies and techniques will allow addressing the ethical issues in ethics-critical decision-making processes. On the whole, this vision will contribute to a more cognizant and responsible use of data, by promoting the ethics-aware development and use of technologies and systems for collecting, storing and processing data.

References

1. Q. Bao, W. K. Cheung, Y. Zhang. Incorporating structural diversity of neighbors in a diffusion model for social networks. *Web Intelligence* 2013: 431–438 (2013)
2. C. Bolchini, E. Quintarelli, L. Tanca. CARVE: Context-aware automatic view definition over relational databases. *Inf. Syst.* 38(1): 45-67 (2013)
3. I. Catallo et al. Top-k diversity queries over bounded regions. *TODS* 38(2): 10:1–10:44 (2013)
4. J. Chomicki, P. Ciaccia, N. Meneghetti. Skyline queries, front and back. *SIGMOD Record* 42(3): 6-18 (2013)
5. A. Caliò *et al.* Topology-driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks. *IEEE TKDE*, To appear
6. P. Ciaccia, D. Martinenghi. Reconciling Skyline and Ranking Queries. *PVLDB* 10(11): 1454-1465 (2017)
7. L. Calderoni, P. Palmieri, D. Maio. Location privacy without mutual trust: The Spatial Bloom Filter *Comput. Commun.*, vol. 68:. 4–16 (2015)
8. L. Calderoni, P. Palmieri, D. Maio. Probabilistic Properties of the Spatial Bloom Filters and Their Relevance to Cryptographic Protocols. *IEEE Transactions on Information Forensics and Security* 13(7): 1710–1721 (2018)

9. X.L. Dong, L. Berti-Equille, D. Srivastava. Data Fusion: Resolving Conflicts from Multiple Sources. *Handbook of Data Quality*. Springer, Berlin, Heidelberg (2013).
10. M. Drosou *et al.*. Diversity in big data: a review. *Big Data* 5(2): 73–84 (2017)
11. S. B. Davidson et al. On provenance and privacy. *ICDT* 2011: 3-10 (2011)
12. G. Koutrika. Data Personalization. In *Data Management in Pervasive Systems*, Springer Verlag, ISBN 978-3-319-20061-3: 213-234 (2015)
13. Y.-H. Fu, C.-Y. Huang, C.-T. Sun. Using global diversity and local topology features to identify influential network spreaders. *Physica A* 433(C):344–355 (2015)
14. L. Floridi, M. Taddeo. What is data ethics? *Phil. Trans. R. Soc.* A374:20160360. http://dx.doi.org/10.1098/rsta.2016.0360 (2016)
15. S. Galhotra *et al.* ASIM: A Scalable Algorithm for Influence Maximization under the Independent Cascade Model. *WWW* 2015: 35–36 (2015)
16. S. Galhotra *et al.* Robust Entity Resolution using Random Graphs. *SIGMOD* (2018)
17. D. Gunning. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency* (DARPA) (2017)
18. Informatics Europe & EUACM. When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making (2018)
19. N. Johnson, J. P. Near, D. Song. Towards Practical Differential Privacy for SQL Queries. *PVLDB*, 11(5): 526–39 (2018)
20. Z.C. Lipton. The mythos of model interpretability. *arXiv*:1606.03490 (2017)
21. J. Mairal et al. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11: 19-60 (2010)
22. K. Mens *et al.* Modeling and managing context-aware systems variability. *IEEE Software*, 34(6): 58–6 (2017)
23. G. Montavon *et al.* Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73: 1-15 (2018)
24. A. Narayanan, S. Vitaly. Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy* (2008)
25. U. Pagallo. On the Principle of Privacy by Design and its Limits: Technology, Ethics and the Rule of Law. *European Data Protection* 2012: 331-346 (2012)
26. P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. *Advances in CryptologyEUROCRYPT* (LNCS), vol. 1592: 223–238 (1999)
27. E. Pitoura *et al.* On Measuring Bias in Online Information. *SIGMOD Record* 46(4): 16-21 (2017)
28. E. Quintarelli, E. Rabosio, L.Tanca. Recommending New Items to Ephemeral Groups Using Contextual User Influence. *RecSys* 2016: 285-292 (2016)
29. T. Rekatsinas et al. SourceSight: Enabling Effective Source Selection. *SIGMOD* 2016: 2157-2160 (2016)
30. J. Stoyanovich, S. Abiteboul, G. Miklau. Data Responsibly: Fairness, Neutrality and Transparency in Data Analysis. *EDBT* 2016: 718-719 (2016)
31. M. Soliman et al. Ranking with uncertain scoring functions: semantics and sensitivity measures. *SIGMOD* 2011: 805-816 (2011)
32. F. Tang *et al.* Diversified social influence maximization. *ASONAM* 2014:455–459 (2014)
33. Y. Tang, X. Xiao, Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. *SIGMOD* 2014: 75–86 (2014)
34. USA ACM. Statement on the Importance of Preserving Personal Privacy - Foundational Privacy Principles and Practices. (2018)
35. J. Wang *et al.* Big data provenance: Challenges, state of the art and opportunities. *Big Data* 2015: 2509-2516 (2015)
36. M. Zehlike *et al.* FA*IR: A Fair Top-k Ranking Algorithm. *CIKM* 2017:1569-1578 (2017)