Lecture Notes in Computer Science 10708

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

More information about this series at http://www.springer.com/series/7409

Riccardo Guidotti · Anna Monreale
Dino Pedreschi · Serge Abiteboul (Eds.)

Personal Analytics and Privacy

An Individual and Collective Perspective

First International Workshop, PAP 2017
Held in Conjunction with ECML PKDD 2017
Skopje, Macedonia, September 18, 2017
Revised Selected Papers

Springer

Editors

Riccardo Guidotti
KDDLab, ISTI-CNR
Pisa
Italy

Anna Monreale
KDDLab, University of Pisa
Pisa
Italy

Dino Pedreschi
KDDLab, University of Pisa
Pisa
Italy

Serge Abiteboul
Inria, École Normale Supérieure
Paris
France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-71969-6 ISBN 978-3-319-71970-2 (eBook)
https://doi.org/10.1007/978-3-319-71970-2

Library of Congress Control Number: 2017960859

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The First International Workshop on Personal Analytics and Privacy (PAP) was held in Skopje, Macedonia, on September 18, 2017. The purpose of the workshop is to encourage principled research that will lead to the advancement of personal data analytics, personal services development, privacy, data protection, and privacy risk assessment with the intent of bringing together researchers and practitioners interested in personal analytics and privacy. The workshop, collocated with the conference ECML/PKDD 2017, sought top-quality submissions addressing important issues related to personal analytics, personal data mining, and privacy in the context where real individual data (spatio temporal data, call details records, tweets, mobility data, transactional data, social networking data, etc.) are used for developing data-driven services, for realizing social studies aimed at understanding nowadays society, and for publication purposes.

The authors were invited to submit original research or position papers proposing novel methods or analyzing existing techniques on novel datasets on any relevant topic including, but are not limited to, the following:

- Personal model summarizing the user's behaviors
- Personal data and knowledge management (databases, software, formats)
- Personal data collection (crawling, storage, compression)
- Personal data integration
- Personal data store and personal information management systems models
- Parameter-free and auto-adaptive methodologies for personal analytics
- Novel indicators measuring personal behavior
- Individual vs. collective models
- Privacy-preserving mining algorithm
- Privacy-preserving individual data sharing
- Privacy risk assessment
- Privacy and anonymity in collective services
- Information (data/patterns) hiding
- Privacy in pervasive/ubiquitous systems
- Security and privacy metrics
- Personal data protection and law enforcement
- Balancing privacy and quality of the service/analysis
- Case study analysis and experiments on real individual data

All submitted papers were reviewed on the basis of technical quality, relevance, significance, and clarity by at least three referees. The Program Committee received 19 submissions and accepted 13 papers. The program of PAP was enriched by the keynote speeches by Bruno Lepri entitled "The Rise of Decentralized Personal Data Markets" and by Serge Abiteboul entitled "Personal Knowledge Management Systems."

We would also like to thank the Program Committee for their work on reviewing the papers. The process of reviewing and selecting papers was significantly simplified through using EasyChair. We thank all attendees to the workshop and hope that this event will enable a good exchange of ideas and generate new collaborations among attendees. The organization of PAP 2017 was supported by the European Community's H2020 Program under the funding scheme "INFRAIA- 1-2014–2015: Research Infrastructures" grant agreement 654024, http://www.sobigdata.eu, "SoBigData: Social Mining & Big Data Ecosystem."

October 2017

Riccardo Guidotti
Anna Monreale
Dino Pedreschi
Serge Abiteboul

Organization

Program Committee

Nicolas Anciaux	Inria, France
Bettina Berendt	KU Leuven, Belgium
Elisa Bertino	Purdue University, USA
Tobias Blanke	King's College London, UK
Francesco Bonchi	ISI Foundation, Italy
Paolo Cintia	ISTI-CNR, Italy
Michele Coscia	Harvard University, USA
Mathieu Cunche	INSA-Lyon/Inria, France
Jon Crowcroft	University of Cambridge, UK
Boxiang Dong	Montclair State University, USA
Wendy Hui Wang	Stevens Institute, USA
Bruno Lepri	MobS Lab at Fondazione Bruno Kessler, Italy
Mirco Musolesi	University College London, UK
Francesca Pratesi	University of Pisa, Italy
Vincenc Torra	IIIA-CSIC, Spain
Jeroen Van Der Hoven	Delft University, The Netherlands
Michele Vescovi	Telecom Italia, Italy

Introduction of the Editors

Personal Analytics and Privacy
An Individual and Collective Perspective

Riccardo Guidotti[1] and Anna Monreale[2]

[1] ISTI-CNR, Via G. Moruzzi, 1, Pisa
Riccardo.Guidotti@isti.cnr.it
[2] University of Pisa, Largo B. Pontecorvo, 3, Pisa
Anna.Monreale@di.unipi.it

1 We All Need to Own and Use Our Own Data

Every year, each person leaves behind her more than 5 GB of *digital breadcrumbs*, disseminated by disparate systems that we use for our daily activities: traveling, communicating, paying for goods, banking, searching the web, listening music, reading, playing, posting or tweeting, screening our health. Five gigabytes, without taking into account photos and videos, otherwise numbers would grow considerably. An avalanche of personal information that, in most cases, gets lost. Only each single individual could connect all this personal information into some personal data repository. No Google or Facebook has a similar power today, and we should very carefully avoid this possibility in the future. The fact that in the contemporary initial phase of a measurable society there are few large harvesters, or "latifundists", who store data on masses of people in large inaccessible repositories in an *organization-centric* model, does not mean that *centralization* is the only possible model, nor the most efficient and sustainable.

Nowadays, data and information belong to big organizations (Amazon, Google, Facebook, etc.) which employ *top-down* control over these data. They can create a mosaic of human behaviors used to extract valuable knowledge for marketing purposes: *our personal data is the new gold*. For example, users produce personal data like Facebook posts, or GPS movements using Google Maps, or online shopping through Amazon, and these data are collected and obscurely employed by these companies for marketing or to produce services. On the other hand, individuals do not have the tools and capabilities to extract useful knowledge from their personal data. This is a *Legrand Star* model [11], i.e., a centralized network model, where users can not directly control and exploit their own personal data. Data owning and usage would require not a *bottom-up* system, but a *Baran Web* model, i.e., a *peer distributed approach*, a network of peers, both individual and companies, in which no single node has absolute control of everything but everyone controls thyself, and has only a partial vision of the surrounding peers. The first brick that must be placed to build this *Web* and to start a change of perspective, is the development of *Personal Data Models*, which are sewn on each individual to fit their subjective behaviors.

Data Mining applied to individual data creates an invaluable opportunity for individuals to improve their *self-awareness*, and to enable *personalized services*. However, nowadays users have a limited capability to exploit their personal data, thus we need a change of perspective towards a *user-centric* model for personal data management: a vision compatible with the data protection reform of EU, and promoted by the World Economic Forum [12, 14, 15].

2 Making Sense of Own Personal Big Data

Although some user-centric models like the *Personal Intelligent Management Systems (PIMS)* and the *Personal Data Store (PDS)* are emerging [1, 5], currently there is still a significant lack in terms of algorithms and models specifically designed to capture the knowledge from individual data and to ensure privacy protection in a user-centric scenario.

Personal data analytics and individual privacy protection are the key elements to leverage nowadays services to a new type of systems. The availability of personal analytics tools able to extract hidden knowledge from individual data while protecting the privacy right can help the society to move from organization-centric systems to user-centric systems, where the user is the owner of her personal data and is able to manage, understand, exploit, control and share her own data and the knowledge deliverable from them in a completely safe way.

Recent works are trying to extend the user-centric models for data management with tools for *Personal Data Analytics* [8]. With *Personal Data Analytics* are indicated the personal data mining processes extracting the user models, and providing self-awareness and personalized services. Personal Data Analytics can be exploited *(i)* to improve the user *self-awareness* thanks to the personal patterns they unveil, and *(ii)* to empower *personalized services* by providing proactive predictions and suggestions on the basis of the user's profile.

In practice, Personal Data Analytics allows a user to make sense of her personal data and to exploit it [9]. Figure 1(a) shows the overall Personal Data Analytics approach. The individual data flow into the PDS and are stored according to one of the possible technique described in the PDS literature [1, 4, 5]. Along the analysis of the continuous digital breadcrumbs, the PDS must consider that it does not exist a unique and constant model describing human behaviors. Indeed, our behaviors will be never "in equilibrium" because we constantly move, we buy new things, we interact with our friends, we listen to music, etc., generating in this way a non-interruptible flow of personal data [2]. Therefore, Personal Data Analytics must be *dynamic* and *adaptable* to continuous changes. The user profile described by Personal Data Analytics can be used to improve the user *self-awareness* and for *personalized services* yet adopting Personal Data Analytics. Self-awareness can be provided for example through a dashboard where the user can navigate and understand her models and patterns. Examples of personalized services can be *recommendation systems* or *predictors* of future actions.

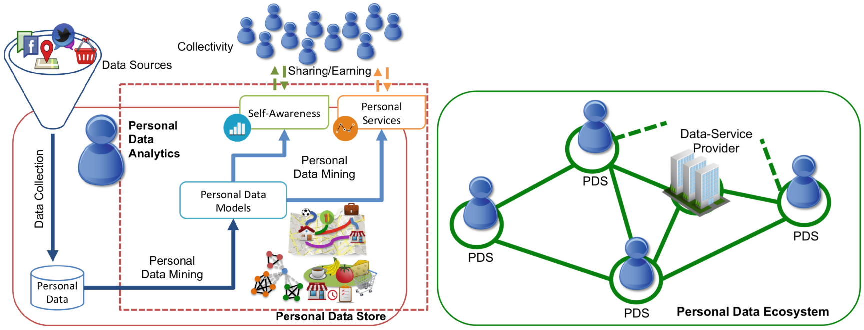

Fig. 1. (a) Personal Data Analytics (left); (b) Personal Data Ecosystem (right).

3 The Personal Data Ecosystem and the Privacy Issues

The PDS extended with PDM offers an augmented image of ourselves, our image reflected in a *digital mirror*. However, passive personal data collection and knowledge mining need to be balanced with *participation*, based on a much greater awareness of the value of own personal data for each one of us and the communities that we inhabit, at all scales.

Personal Data Analytics enables the comparison of our individual patterns with the collective ones of the communities we belong to, provided we have a way to interact and collaborate with other peers, individuals and institutions that are, in turn, equipped with their PDS's and connected to each other in a network. Figure 1(a) (top right) shows how, to provide and obtain improved *self-awareness* and *personalized services*, a user can share information and, at the same time, earn knowledge, by communicating with the *collectivity*.

This enables a *Personal Data Ecosystem (PDE)* (Fig. 1(b)) that can be modeled as distributed network of peers, both individual users and public or private institutions and companies, each one with their own PDS and PDM [8]. The PDE can generate an innovative form of *collective awareness*, characterized by a self-reinforcing loop where *(i)* superior individual knowledge is created by comparison with collective knowledge, enhancing individual ability to better align own goals with common interest, and *(ii)* superior collective knowledge is created through the active participation of individuals in a decentralized system, i.e., without centralization of unnecessarily large amounts of information. The PDE is in line with the *peer progressive* idea of a decentralized network where news, money, and knowledge come from the periphery instead of from the center [11]. In [10] it is compared the concept of "wise king" (centralized system) against the Adam Smith's "invisible hand" regulating a decentralized self-organizing system. Furthermore, the PDE idea outlines the *Nervousnet* project: a globally distributed, self-organizing, techno-social system for answering analytical questions about the status of world-wide society, based on social sensing, mining and the idea of trust networks and privacy-aware social mining [7].

Although the PDE setting enables new opportunities for individuals who may exploit their own data to improve the daily life with more and more self-awareness, it is impossible to ignore the possible privacy risks that may derive from the sharing of personal data and the knowledge extractable from them. Indeed, the worrying aspect of this story is that often, individual data provide a very fine detail of the individual activities and thus, in case of *sensitive* activities the opportunities of discovering knowledge increase with the risks of *privacy violation*. The threat goes as far to recognize personal or even sensitive aspects of their lives, such as home location, habits and religious or political convictions. Managing this kind of data is a very complex task, and often we cannot solely rely on de-identification (i.e., removing the direct identifiers contained in the data) to preserve the privacy of the people involved. In fact, many examples of re-identification from supposedly anonymous data have been reported in the scientific literature and in the media, from health records to GPS trajectories and, even, from movie ratings of on-demands services. Several techniques have been proposed to develop technological frameworks for countering privacy violations, without losing the benefits of big data analytics technology [6]. Unfortunately, most of the research work done in the context of privacy-preserving data mining and data analytics focuses on an organization-centric model for the personal data management, where the individuals have a very limited possibility to control their own data and to take advantage of them according to their needs and wills. PDE instead provides to individuals the chance to have a central and active role in the control of the lifecycle of her own personal data introducing also a layer of transparency. In particular, it enables individuals to control a copy of their data and/or the knowledge extracted from them. In practices, the individuals acquire the right to dispose or distribute their own individual information with the desired privacy level in order to receive services or other benefits or in order to increase their knowledge about themselves or about the society they live in. In this setting to guarantee the information sharing with the level of desired privacy level, *Privacy-by-Design* data transformations [3, 13] must be applied before data leave the user. This guarantees the prevention of privacy attacks to the PDE addressing possible privacy issues with a pro-active approach. This encourages the voluntary participation of users, limiting the fear and skepticism that often leads people to not access the benefits of extracting knowledge from their own data, both at personal and collective level.

Acknowledgments. This work and workshop has been partially supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014–2015: Research Infrastructures" grant agreement 654024, "SoBigData: Social Mining & Big Data Ecosystem" (http://www.sobigdata.eu).

References

1. Abiteboul, S., André, B., Kaplan, D.: Managing your digital life. Commun. ACM, **58**(5), 32–35 (2015)
2. Ball: Why Society is a Complex Matter. Springer (2012)
3. Cavoukian: Privacy design principles for an integrated justice system. TR (2000)

4. Vescovi, M., et al.: My data store: toward user awareness and control on personal data. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 179–182 (2014)
5. de Montjoye, Y.-A., Shmueli, E., Wang, S.S., Pentland, A.S.: openpds: protecting the privacy of metadata through safeanswers. PloS one **9**(7), e98790 (2014)
6. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 14:1–14:53 (2010)
7. Giannotti, et al.: A planetary nervous system for social mining and collective awareness. Eur. Phys. J. Spec. Top. **214**(1), 49–75 (2012)
8. Guidotti, R.: Personal data analytics: capturing human behavior to improve self-awareness and personal services through individual and collective knowledge (2017)
9. Guidotti, R., et al.: Towards user-centric data management: individual mobility analytics for collective services. In: ACM SIGSPATIAL, pp. 80–83 (2015)
10. Helbing: The automation of society is next: how to survive the digital revolution. Available at SSRN 2694312 (2015)
11. Johnson: Future Perfect: The Case for Progress in a Networked Age. Penguin (2012)
12. Kalapesi: Unlocking the value of personal data: from collection to usage. In: WEF (2013)
13. Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., Pedreschi, D.: Privacy-by-design in big data analytics and social mining. EPJ Data Sci. **3**(1), 10 (2014)
14. Pentland, A., et al.: Personal data: the emergence of a new asset class. In: WEF (2011)
15. Rose, et al.: Rethinking personal data: strengthening trust. In: WEF (2012)

Abstracts of Invited Talks

The Emergence of Personal Data Markets

Bruno Lepri

Fondazione Bruno Kessler, 38123 Trento, Italy

Abstract. The almost universal adoption of mobile phones, the exponential increase in the use of Internet services and social media platforms, and the proliferation of wearable devices and connected objects have resulted in a massive collection of personal data that represents an invaluable resource for designing and building systems able to understand the needs and activities of both individuals and communities - so as to provide personalized, context-aware services. Hence, many public services of societal interest (e.g. emergency response, city planning, etc.) are dependent on this data. At the same time, many commercial services require sensitive data such as location, purchase, or browsing history. However, this scenario raises unprecedented privacy challenges derived from the collection, storage and usage of vast amounts of personal data. The core problem is how can an individual share sensitive data for a public service or a desired commercial service and be sure that the data will only be used for the intended purpose? This question implicitly recognizes the risks in terms not only of possible abuses in the usage of the personal data but also of the "missed chance for innovation" that is inherent to the current dominant paradigm of siloed data collection, management, and exploitation, which precludes participation to a wide range of actors, most notably the very producers of personal data (i.e. the users). Recently, new user-centric models for personal data management have been proposed, in order to provide individuals with more control of their own data's life-cycle. To this end, researchers and companies are developing repositories which implement medium-grained access control to different kinds of personally identifiable information (PII). Previous work has also introduced the concept of personal data markets in which individuals sell their own personal data to entities interested in buying it. While personal data markets might be the next natural step in designing technology to support a transparent and privacy-preserving use of personal data, they are still at a research stage due to a number of technological and regulatory challenges. In order to implement a market, it is necessary to connect potential buyers (demand) with sellers (offer) and provide a trustworthy mechanism for the exchange of goods or value between buyers and sellers. In a personal data market, the sellers are individuals (who own their personal data), the buyers are corporations, researchers, governments, etc., and the mechanism for the exchange of "goods" is still to be defined. In my talk, I provide a possible answer to such issue describing recent efforts to develop a user-centric Personal Data Market approach, coupled with cryptographic guarantees on data access and usage.

Personal Knowledge Base Systems

Serge Abiteboul and David Montoya

Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL
Research University, 75005 Paris, France

The typical Internet user has personal data spread over several devices and across several online systems. Geeks already know how to control their personal data. It is now (or soon will be) possible for everyone to do the same, and there are many advantages to doing so. Everyone should now manage his/her personal info management system. We explain the main features of *personal information management systems* (sometimes called, self-data management systems). We consider advantages they provide from societal viewpoints. We argue that they are coming because of a number of reasons, political, economical, technical, and societal.

We believe that a main argument for their future success is that they enable new functionalities based on knowledge integration, and thus that the future systems are personal *knowledge* management systems. Such a system should integrate into a coherent whole data of very different nature, in particular, emails and other messages, calendars, contacts, GPS locations, web searches, bills, bank and medical records, etc. The integration of data that have typically to be exported from various systems should be performed by transforming all the information into a single knowledge base on a machine the user controls. The resulting system thus acts as a digital home for all the digital knowledge of the person.

This integration allows the user to query her personal information within and across different dimensions. It also enables performing analytics to learn from all this information. We believe that many users would be reluctant to give access to all their information to a single company, or if they do that, they will require strong guarantee of privacy. The fact that the integration, the query evaluation and analysis happens on a system directly controlled by the user guarantees her privacy.

To illustrate the possibly very rich derivation of knowledge, suppose that a user, say Alice, met a friend, say Bob. The agenda indicates the time of a meeting with "Bob". Some text messages may remove the ambiguity on which Bob she met (Alice may know several Bobs). The phone number he used has long been aligned wih a specific Bob. The location of Alice, as provided by her phone GPS, provides the name of the bar where they met. Bob' timeline in his social network provides pictures of the event. Finally, from her search history, we can even learn about topics they discussed. A meeting that had traces in different systems turned into a rich event in the KB.

Designing such a personal KB is not easy: Data of completely different nature has to be modeled in a uniform manner, pulled into the knowledge base, and integrated with other data. Different from related work in the area of information integration, we cannot rely on outside sources, we cannot expect redundancy, and we have a very limited and particular set of attributes.

We will mention such a system, the Thymeflow system that for instance aligns events based on time, e.g., a meeting in calendar and a GPS position, or on space, an address in contacts and a GPS position.

Acknowledgments. I will discuss works performed with collegues that I would like to thank:

- Benjamin André, Cozy Cloud
- Daniel Kaplan, Fing
- Amélie Marian, Rudgers University
- Thomas Pellissier Tanon, ENS Lyon
- Pierre Senellart, ENS
- Fabian M. Suchanek, Telecom ParisTech

The bibliography provides references to these works. More references can of course be found in these papers.

References

1. Abiteboul, S., Andr, B., Kaplan, D.: Managing your digital life. Commun. ACM **58**(5), 32–35 (2015)
2. Abiteboul, S., Marian, A.: Personal information management systems. In: Tutorial, EDBT/ICDT Conference (2015). https://www.slideshare.net/ameliemarian
3. Montoya, D., Tanon, T.P., Abiteboul, S., Suchanek, F.M.: Thymeflow, a personal knowledge base with spatio-temporal data. In: CIKM 2016, pp. 2477–2480 (2015)
4. Montoya, D., Abiteboul, S., Senellart, P.: Hup-me: inferring and reconciling a timeline of user activity from rich smartphone data. In: SIGSPATIAL/GIS 2015, pp. 62:1–62:4 (2015)

Contents

Privacy on Personal Data Based Applications

Personal Analytics: An Individual and Collective Perspective