An Analysis of the Consequences of the General Data Protection Regulation on Social Network Research

ANDREAS KOTSIOS, Uppsala University
MATTEO MAGNANI and DAVIDE VEGA, Uppsala University
LUCA ROSSI and IRINA SHKLOVSKI, IT University Copenhagen

This article examines the principles outlined in the General Data Protection Regulation in the context of social network data. We provide both a practical guide to General Data Protection Regulation–compliant social network data processing, covering aspects such as data collection, consent, anonymization, and data analysis, and a broader discussion of the problems emerging when the general principles on which the regulation is based are instantiated to this research area.

CCS Concepts: • **Applied computing** → **Sociology**; • **Social and professional topics** → *Privacy policies*;

Additional Key Words and Phrases: General Data Protection Regulation, GDPR, social network analysis, social media

ACM Reference format:
Andreas Kotsios, Matteo Magnani, Davide Vega, Luca Rossi, and Irina Shklovski. 2019. An Analysis of the Consequences of the General Data Protection Regulation on Social Network Research. *ACM Trans. Soc. Comput.* 2, 3, Article 12 (November 2019), 22 pages.
https://doi.org/10.1145/3365524

12

1 INTRODUCTION

In the past decade, online social network platforms have become a major source of data to study human and social behavior [3, 14]. The availability of persistent and searchable traces of human communication on a large scale [6] provided new, previously inconceivable opportunities for unobtrusive research but also raised new questions related to the potential misuse of personal information [17, 24]. Following events such as the Cambridge Analytica scandal, and related restrictions on research-related data access established by large social media companies, some Internet researchers have highlighted the necessity and complexity of ensuring that "independent, critical research in the public interest can be conducted while protecting ordinary users' privacy".[1] For

[1] https://medium.com/@Snurb/facebook-research-data-18662cf2cacb.

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 727040 (Virt-EU).

Q1

Authors' addresses: A. Kotsios, Faculty of Law, Uppsala University, Uppsala, Sweden; email: andreas.kotsios@jur.uu.se; M. Magnani and D. Vega, InfoLab, Department of Information Technology, Uppsala University, Uppsala, Sweden; emails: {matteo.magnani, davide.vega}@itt.uu.se; L. Rossi and I. Shklovski, IT University Copenhagen, Copenhagen, Denmark; emails: {lucr, irsh}@itu.dk.

2469-7818/2019/11-ART12 $15.00
https://doi.org/10.1145/3365524

European researchers, this context is further complicated by the recent advent of the General Data Protection Regulation[2] (GDPR), which came into force on May 25, 2018.

The GDPR is a piece of European legislation regulating how natural persons should be protected with regard to the processing of their personal data. The GDPR applies to processing of personal data in very similar ways in all European Union (EU) Member States, for all sectors (public or private) and all purposes (commercial and non-commercial). This includes research performed by private companies or public universities and other research institutions. The regulation has been welcomed as a progressive step toward rectifying the glaring power imbalance in current mass digital data collection by entities that develop, maintain, and control access to digital infrastructures. The GDPR has two main goals.[3] The first goal is to protect the fundamental rights and freedoms of the data subjects by creating a protective regiment with regard to the processing of personal data.[4] This is because new technologies and organizational models both in the private and public sector have made it easy to gather, use, combine, aggregate, or otherwise process a vast amount of personal data without sufficient controls or oversight. The second goal is to create the optimal conditions so that the free flow of personal data—in parallel to the free movement of goods and services—can take place within the EU, supporting the creation of the European Single Market. The GDPR is intended to provide a way of achieving the free flow of data within the EU while ensuring protection of the fundamental rights and freedoms for individuals.

The regulation replaces the earlier Data Protection Directive,[5] which, as a directive, was adopted and implemented through different national laws by every EU Member State, resulting in at times a confusing patchwork of national regulations. More importantly, the earlier directive did not have any specific focus on research. Rather, the main regulatory mechanisms were codes of conduct and ethical guidelines advocating good practices but rarely systematically codifying these. In contrast, the GDPR explicitly recognizes the particularities of data processing in research through a series of formally specified research exemptions, which have important consequences on the feasibility and lawfulness of social network research projects in practice. This includes the ability to limit and even avoid restrictions on secondary processing and the processing of sensitive categories of data,[6] to override the subjects' right to object to processing and erasure as long as relevant safeguards are implemented,[7] and to collect some types of data without consent for some types of processing.[8]

The impact of the new rules on the practice of research is unclear, and this is especially relevant to researchers studying social network data (both on- and offline), where, for example, the subjects participating in a study may provide information about non-participants and the collected data is more difficult to effectively anonymize than in other research fields. This article considers the impact and implications of the GDPR and of the research exemptions built into the law on the activities of researchers engaging in social network analysis in general and in the specific case of online social networks.

[2]Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.
[3]Rec 1-7 GDPR and SOU 2017:50, p78.
[4]In the context of GDPR, "personal data" means any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, particularly by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.
[5]Directive 95/46/EC.
[6]Art 6(4) and rec 50 GDPR.
[7]Art 89 GDPR.
[8]Art 6(1)(e) and (f); rec 47 and 157 GDPR.

The text of the GDPR is complex and not specifically targeted to researchers, but its content will impact research practices in significant ways that depend on the specific research field. However, universities and other research institutions are of course providing general information about the GDPR to their employees and have a data protection officer who can be contacted for specific mapters. Yet, GDPR-compliant strategies need to be instantiated to the specific research problem, and many of these are difficult to interpret without domain-specific knowledge. Therefore, researchers should themselves be aware of the implications of the GDPR and have reflected about it. Several works have already been published to address this issue and to examine the impact of the GDPR on research in general [11, 22, 26] and on specific research fields [25]. However, none have addressed the implications of the GDPR for social network analysis, where data processing differs from other quantitative approaches. For example, Borgatti and Molina [4] point out that respondent anonymity is not an option if we want to know who is talking about whom, which is necessary to define edges in the network. In addition, subjects providing information about their social relations may generate data about individuals not included in the study: a participant mentioning that she often performs some activity with someone may reveal a lot about this other person depending on the type of activity. Another issue is the fact that in social networks, it is often possible to identify specific roles based on the network structure, with a limited number of individuals in each role, examples being high-degree and high-betweenness[9] nodes, as well as other special network configurations. Once these few nodes have been identified, it becomes very simple to connect them to specific individuals using some basic knowledge of the studied organization. For this reason, network data is often impossible to fully anonymize. In this article, we will also identify specific issues related to data protection emerging when social network analysis is applied to contexts such as the analysis of large-scale networks of social relations derived from social media data.

In the next section, we present an overview of the GDPR, including the terminology used in the rest of the article. This short section is necessary to make this article self-contained, and it can be skipped by the reader who is already familiar with the main actors, concepts, and principles introduced by the regulation. The following section is organized along the main steps and problems of a typical social network research process. We start by discussing approaches to data collection, also highlighting the differences between data collected directly from the data subjects or indirectly, such as through social media application programming interfaces (APIs). We also discuss topics such as consent, data anonymization, profiling, and storage of the networks. We conclude the article with more general considerations about the implications of the GDPR for commercial data controllers, as well as for the future of network data repositories that represent important teaching and training tools. We also suggest that the GDPR should be a new important element to be considered in the ongoing discussion about the establishment of a code of conduct for social network research.[10]

Please note that our objectives are (1) to provide a general understanding of the impact of the GDPR on social network research for scholars with no background in law, under the assumption that their institutions will be able to fill in the details about local regulations but may be unaware

[9]Degree and betweenness are so-called centrality measures that can be used to identify important actors in a network.

[10]The GDPR applies only to the processing of personal data by entities established in the EU regardless the place of processing, or in general to processing of personal data of data subjects who are in the EU, as long as the processing is related to the offering of goods and services in the Union or the monitoring takes place in the Union. Our presentation will often take the perspective of a European public university, and we will extend the discussion to other cases regulated by this law when relevant. However, the principles defined in the GDPR are worthy of consideration even for researchers outside the EU processing data from non-EU subjects, as these principles highlight general fundamental issues to be considered when processing personal data.

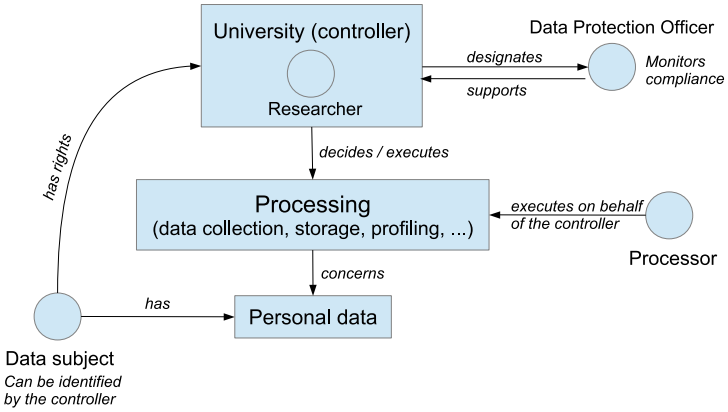

Fig. 1. Typical configuration of the GDPR ecosystem.

of the aspects more specific to social network data, and (2) to highlight some controversial issues. It is not our goal to provide a detailed technical legal analysis of the GDPR; for this, we refer to the relevant current debate and scholarly work.[11] Similarly, we do not provide legal analyses of specific cases, as this is outside the scope of this article.

2 THE GDPR ECOSYSTEM: OVERVIEW AND TERMINOLOGY

Before we proceed to our analysis of how the GDPR may affect social network research in practice, it is important to make clear the fundamental terms and ideas of this piece of law. The regulation is complex in terms of length (88 pages, 173 recitals, 99 articles), breadth of coverage, and depth. Here, we present the GDPR concepts and principles that we find are the most relevant for social network analysis research. In the next section, we will discuss the (often unclear) role that these concepts and principles can play in the various phases of a social network analysis process. In the article, we will refer to articles in both the GDPR and the recitals. Even though recitals are not part of the operative text of the regulation, their role is of great importance since their purpose is "to set out concise reasons for the chief provisions of the enacting terms".[12] We will not discuss here the relationship between recitals and articles, but it is worth keeping in mind that when the legal text of the regulation is somewhat ambiguous, it will normally be interpreted in light of the relevant recital.[13]

Figure 1 exemplifies the main concepts described in this section in the context of a typical academic research project.[14] The data subject is defined as every individual (natural person) who is identified or may be identified by the controller or third parties, directly or indirectly, by the act of processing her personal data. The natural or legal persons who decide how and why the personal

[11] See, for example, L. Feiler, N. Forgó, and M. Weigl, *The EU General Data Protection Regulation (GDPR): A Commentary*. (Globe Law and Business 2018) and C. Kuner, L. Bygrave, and C. Docksey (Eds.), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2019).

[12] Legal Service, European Parliament et al., *Joint Practical Guide of the European Parliament, the Council, and the Commission for Persons Involved in the Drafting of Legislation Within the Community Institutions* (2015), section 10, available at https://publications.europa.eu/en/publication-detail/-/publication/3879747d-7a3c-411b-a3a0-55c14e2ba732.

[13] Tadas Klimas and Jurate Vaiciukaite, *The Law of Recitals in European Community Legislation* (*ILSA Journal of International & Comparative Law* Vol. 15, 2008).

[14] In Section 5.4, we discuss the difference between universities and commercial research.

data will be processed are the data controllers, whereas the ones who process the data on behalf of the controller are the data processors.[15]

What constitutes personal data is defined quite broadly as any information that does or may lead to the identification of a natural person.[16] The term *processing* is defined similarly broadly as "any operation or set of operations on personal data or sets of personal data",[17] including data collection. A special type of processing is profiling. With an equally broad definition, the term *profiling* is defined as "any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements."[18]

A data protection officer is the person designated by the controller and/or the processor in case the processing (1) is carried out by a public authority, (2) contains systematic monitoring of data subjects, or (3) consists of large- scale processing of special categories of data.[19]

These definitions, when applied to research, make clear that any research-based processing of social network data that not only directly identify but also possibly may identify (by the same researchers or third parties) individuals will be regulated under the GDPR.

Typically, in cases where research is conducted under the auspices of a university, the university is considered to be the data controller. Although universities are supposed to have organizational measures with regard to the GDPR, the researchers, as employees of the university, who one way or another process personal data as part of their role, are also expected to have an understanding of the GDPR as they design data collection and analysis protocols.

It may happen that other entities assist with the processing but do not decide the purposes and manner of it. For example, a researcher may pay individuals who are not employees of the university to perform a data collection. Where these entities are only following the guidelines of the controller, then these external parties can be seen as data processors, where the "processing by a processor shall be governed by a contract or other legal act."[20] If the research is conducted by a private actor, such as a company, then it is the company that is the controller, and possible external sub-contractors (including researchers) may constitute the processors. It can also be so that a private party and a university can both jointly be regarded as controllers, depending on what agreement exists between these two parties.

As we can see here, in case of more than one entity—meaning different legal entities—being involved in the processing of personal data, determining the controller(s) is not always an easy task. It is not sufficient that one entity processes data on behalf of another entity, as it is possible that it also processes these data for its own purposes.[21] It is not sufficient either that a contract may explicitly state the roles of the entities, as they in reality may act in a different way.[22] The capacity of the controller is based on factual elements and circumstances, on whether or not an entity can—and does—indeed determine the *purposes* and the *means* of processing: the "whys" and the "hows." Some questions that help in determining the role of controller are the following: why

[15] Art 4(8) and 4(9) GDPR.
[16] Art 4(1) GDPR.
[17] Art 4(2) GDPR.
[18] https://gdpr-info.eu/recitals/no-71/.
[19] Art 37 GDPR.
[20] Art 28(3) GDPR.
[21] Art 29 Data Protection Working Party (2010) Opinion 1/2010 on the concepts of "controller" and "processor," WP 169, p9.
[22] Ibid.

Table 1. Seven Basic Principles in the GDPR

P1:	Lawfulness, fairness, and transparency
P2:	Purpose limitation
P3:	Data minimization
P4:	Accuracy
P5:	Storage limitation
P6:	Integrity and confidentiality
P7:	Accountability

is this processing taking place, who initiated it, would an entity process the data if not asked by another entity, and if so under what conditions?[23]

It is important to keep in mind that the *purpose* of the processing can be defined only by the controller, meaning that if there is more than one entity defining the purposes of the processing, then these entities are joint controllers. The *means*, however, namely the decision on organizational and technical matters, can be delegated by the controller to a processor. However, "substantial questions which are essential to the core of lawfulness of processing are reserved to the controller."[24] If an entity has the power to decide, for example, issues related to the period of storage or access privileges, this entity is, then, de facto a controller concerning this part of the use of data.

These complicated distinctions are important to consider and discuss with relevant internal data protection officers because their particular specifications can have an impact on the obligations of the researcher or, alternatively said, on how to comply with the GDPR. As a simple example, it is the controller who is responsible for providing specific information to the data subjects.

Given the diversity of research approaches, it is important that researchers understand the particular aspects of the regulation that apply to them. This also means that any collaborative research project must consider what institutional agreements must be made with respect to data processing: a process that may take additional time and must be planned for.

The regulation introduces seven important principles, listed in Table 1, to be followed when processing personal data.[25] (P1) The data must be processed in a lawful, fair, and transparent way. (P2) Personal data may only be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes. (P3) The data may be processed only if they are adequate, relevant, and limited to what is necessary with regard to the purpose of processing. (P4) Only data that are accurate and up to date, to the level that it is possible, may be processed. (P5) Personal data may only be processed for a period that is necessary for the processing, and therefore the controllers must create criteria to determine what retention periods are suitable for their purposes. (P6) The controllers must apply technical and organizational measures to protect personal data they control against unauthorized and unlawful processing, as well as accidental loss, destruction, or damage. (P7) The data controllers have the responsibility to be compliant and to be able to demonstrate compliance when needed, which implies that written records must be kept on whether and how the controller is compliant.[26] These principles have implications for social network analysis research, which are detailed in the following.

[23] Art 29 Data Protection Working Party (2010) Opinion 1/2010 on the concepts of "controller" and "processor," WP 169, p8.
[24] Art 29 Data Protection Working Party (2010) Opinion 1/2010 on the concepts of "controller" and "processor," WP 169, p15.
[25] Art 5 GDPR.
[26] With regard to the principle of accountability, we would also like to draw attention to the provisions of art 24 et seq GDPR defining the liabilities, responsibilities, and general obligations of the controllers and the processors.

Table 2. Summary of General Rules and Exemptions to Be Considered During the Social Network Analysis Process

	General Rule	Exemption	Details
1	Identify the roles w.r.t. the GDPR ecosystem (data subjects, controllers, processors, DPO, etc.) and the data flows.	No	This can be challenging in some cases; consult the DPO if uncertain.
2	Identify the nature of the data (personal/non-personal/sensitive).	No	In case of sensitive data, we can process it (1) if we have explicit consent; (2) if the data were manifestly made public by the data subject (use this carefully); or (3) in case of research purposes, if there are suitable safeguards (e.g., pseudonymization, approval from an ethics committee).
3	Identify explicit and legitimate purposes for the processing.	Yes	The specification in case of research can be a bit more general (e.g., the general research area or part of the project, not specific analytical tasks). Some specification of the intended purpose, however, is necessary.
4	Identify the lawful basis for data processing.	No	Based on national legislation, which is still being produced, some actors conducting research (e.g., universities) might be assumed to operate in the public interest, and therefore the public task basis may primarily be used. Otherwise, the consent and legitimate interests bases should be examined.
5	Define clear temporal limits for data processing. Non-anonymized data can be kept for no longer than is necessary for the purposes of the processing.	Yes	More extended periods may apply in case of research as long as appropriate safeguards are implemented.
6	Put in place technical and organizational measures to protect the data. For example, ensure privacy by design, and by default, pseudonymize the data as soon as possible.	No	The measures should be proportionate to the aim pursued.
7	In case of profiling, perform a DPIA.	No	Consider with the DPO whether a DPIA is necessary.

Note: *Exemption* column indicates whether explicit exemptions exist for research, and exemptions (if any) and other considerations are indicated under the *Details* column. DPO, data protection officer; DPIA, data protection impact assessment. (Part 1)

3 THE GDPR IN THE SOCIAL NETWORK ANALYSIS PROCESS

The principles mentioned in the previous paragraph and in Table 1 need to be instantiated to the specific cases. In this section, we will discuss what implications data processing in the context of research has on the practical enactment of the principles. We will also detail the meaning of these principles when they regulate the processing of social network data, emphasizing the cases where ambiguities arise.

A summary of the main GDPR-related aspects that should be considered during a social network analysis process, including a list of exemptions that can be applied in research, is presented in Tables 2 and 3. Detailed explanations are presented in the text.

Table 3. Summary of General Rules and Exemptions to Be Considered During the Social Network Analysis Process

	General Rule	Exemption	Details
8	Inform the data subjects about the collection, purposes, and their rights at the time the data are obtained (if obtained directly from the data subject) or within a reasonable period after the data are obtained and no later than a month (if the data are obtained indirectly).	Yes	For secondary data, providing information is not necessary if the provision of such information proves impossible or would involve a disproportionate effort, if this is likely to render impossible or seriously impair the achievement of the objectives of the processing.
9	Collect only adequate, relevant, and limited data to what is necessary to achieve the purposes of the processing.	Yes	As the purpose may be specified in less precise terms (see the exception to Rule 3), this rule is also affected. Consider deleting unwanted data as soon as possible, acknowledging and documenting the process.
10	Data subjects have the right to check if there is data concerning them and the right to obtain these data.	No	Even if not part of the GDPR, national laws may still restrict this right (e.g., secrecy acts).
11	Data subjects have the right to have the data concerning them erased.	Yes	Not necessary if it is likely to render impossible or seriously impair the achievements of the objectives of the processing. National laws may also restrict this right.
12	Keep data accurate and up to date.	No	
13	If a new purpose emerges, new legal bases for data processing should be identified.	Yes	If the new purpose is research, further processing is considered to be compatible to the initial purpose.
14	If the controller changes the purpose of the processing, information must be provided to the data subject prior to this processing.	Yes	See the exception to Rule 3 about the increased flexibility in the specification of the purpose in case of research.
15	Keep written records to demonstrate compliance.	No	

Note: The *Exemption* column indicates whether explicit exemptions exist for research, and exemptions (if any) and other considerations are indicated under the *Details* column. (Part 2)

3.1 Lawful Bases for Data Processing

The first basic principle of GDPR states that the data must be processed in a lawful, fair, and transparent way. This means that for data to be processed, there has to be some lawful basis for doing so. The GDPR lists six lawful bases for processing of personal data[27]: (1) the data subject has given her consent, (2) it is necessary for the performance of a contract, (3) it is necessary for the controller to comply with a legal obligation, (4) it is necessary to protect individuals' (the data subject's and/or other natural persons') vital interests, (5) it is necessary for the performance of a task carried out in the public interest, and (6) it is necessary for the purposes of legitimate interests pursued by the controller as long as these interests are not overridden by interests and fundamental rights and freedoms of the data subjects. Even though there are no specific lawful bases that are *a priori* dedicated to research, the three most relevant tend to be the consent of the data subject (1), the task carried out in the public interest (5), and the legitimate interests of the controller (6).

With a long history starting in medical science, the practice of informed consent has for long time been the central pillar of research practices involving human subjects [9]. A key element of

[27] Art 6 GDPR.

the GDPR is that, addressing a growing lack of satisfaction toward the efficacy of informed consent practices [21], it provides well-defined research exemptions.

By examining the GDPR closer, we can notice that when it comes to the question of which lawful basis should be used when processing personal data in general, the most important parameters to take into consideration are the identity of the controller,[28] the purposes of processing, and the context of processing. Depending on these parameters, the controller must decide which lawful basis to use for processing. In the case of research, the following lawful bases seem to be the most relevant: the data subject has given her consent for the processing of her personal data,[29] the processing is necessary for the performance of a public task,[30] and/or it is necessary for the purposes of the legitimate interests pursued by the controller.[31]

In the case where a controller is a university, it may be most suitable to use as a lawful basis that the processing is necessary "for the performance of a task carried out in the public interest."[32] The definition of such tasks is left to Union or Member States law.[33] There is, however, no need for an explicit statutory provision as long as there is a clear basis in law.[34] Even in cases where no national legislation is introduced with regard to it, it should be accepted that pubic actors, such as universities, may use this lawful basis for processing of personal data.[35] Considering that in many countries universities—often even private ones—are considered to be public authorities by law and they act on carrying out tasks of public interest, such as conducting research,[36] the public task basis for processing personal data seems to be the appropriate lawful basis for a social network research project, as long as the processing is necessary for that project.[37] This lawful basis puts the onus of ensuring that the rights of the data subject are balanced against the public interest goals of institutions, whose aims presumably are oriented toward the greater good. This basis is not available at all to commercial organizations and research labs—at least as long as no law provides for that—which must rely on consent or the legitimate interest basis to process personal data.

With regard to the use of consent[38] as a lawful basis for the processing of data in research, there are some things that have to be taken into consideration. The first one is that even though this lawful basis can also be used for the processing of personal data by a research project, an entity may use this lawful basis only "if a data subject is offered control and is offered a genuine choice with regard to accepting or declining the terms offered or declining them without detriment."[39] If this is not possible, something that in social network research—and research in general—can be the case, then this lawful basis should not be used.[40] Additionally, for public universities, since they

[28]Even though the GDPR applies both to public and private actors, the identity of the controller may lead to different outcomes, as we will illustrate later in this article.
[29]Art 6.1(a) GDPR.
[30]Art 6.1(e) GDPR.
[31]Art 6.1(f) GDPR.
[32]Art 6.1(e) GDPR. See also SOU 2017:50.
[33]Art 6.3 GDPR.
[34]Rec 41 GDPR.
[35]SOU 2017:50, p18.
[36]See, for example, in the UK the Freedom of information Act 2000 and in Sweden the Higher Education Act 1992:1434.
[37]According to art 6.2 and 6.3 GDPR as well as rec 45 GDPR, it is stated that Union or Member State law shall define whether the controller performing a task of public interest can be a legal person governed by public law or by private law.
[38]It is not the goal of this work to make an analysis on consent as a lawful basis in general. For a better understanding, we refer to the Article 29 Working Party Guidelines on consent under Regulation 2016/679, but it is worth reminding here that a consent for processing of personal data by a data subject has to be freely given, specific, informed, and unambiguous.
[39]Article 29 Working Party Guidelines on consent under Regulation 2016/679, p3.
[40]Here, it is also important to consider that, as we will argue later in this article, it can be difficult to provide information to the data subjects of a network research project, and therefore it can similarly be challenging to provide the possibility for an informed consent.

are public authorities, researchers must always assess whether or not the consent provided by the data subjects is valid, namely if it is indeed freely given or it is given as a product of imbalance in powers between the university and the data subjects.[41] Last, one should make a distinction regarding the term *consent* as developed in the GDPR and as an "ethical standard and procedural obligation."[42] That means that it can be so that the lawful basis for processing is the public task basis, art 6.1(e) GDPR, but consent is used as an additional safeguard. In this case, it is not two lawful bases used for the processing of personal data but only one, the public task base; consent is only a procedural obligation and not the lawful basis provided for in art 6.1(a).

The third possible lawful basis for research is that the processing is necessary for the legitimate interests[43] pursued by the controller or a third party. In general, this basis is the most flexible one; at the same time, a controller should be very careful when using it as a lawful basis. More specifically, the controller should prove that there is some legitimate interest; that there is a necessity to process personal data for this legitimate interest; that the interests and rights of the data subjects are not violated, namely that there is a minimal privacy impact; and that the data subject would not be surprised by such a processing or is not likely to object. An important thing to remember here is that it cannot be used as a basis in cases where public authorities are processing personal data in the performance of their tasks.[44] Therefore, a public university processing data in the performance of their tasks, which also include research activities, should probably avoid basing the processing conducted for a research project on the legitimate interest basis.

One last thing that we would like to add here is that if the personal data processed are of sensitive character, an entity conducting research—at least an entity, such as a university, that bases its research activities on some piece of legislation—may primarily base the lawful processing of such data on the fact that the processing is necessary for scientific research purposes as long as appropriate measures are deployed according to art 89.1 and the research is based on a law "which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject" according to art 9.2(j) GDPR.[45] Following the same argumentation as earlier, we could, however, claim that if the processing is not necessary or if there is still no specific legislation with regard to processing for research purposes, consent could also be used as a lawful ground for such processing, according to art 9.2(a) GDPR.[46]

3.2 Data Collection

Social networks can be obtained through a wide range of data collection strategies. In the following, we detail different approaches to data collection for social network analysis and consider the corresponding consequences of the GDPR. It is worth noting that we focus on networks where nodes represent natural persons: the GDPR does not apply when nodes represent companies, or

[41] In most research projects, this should not be a great issue because data subjects in a network research project do not normally have a direct connection to a university, but it is still worth considering possible problems that may arise.
[42] Art 29 Working Party Guidelines on consent under Regulation 2016/679, p28.
[43] The meaning of legitimate interests is to be interpreted widely and contain both trivial and more important interests, commercial or societal, among others.
[44] Art 6.1 para 2 GDPR.
[45] Art 9.2(g), namely that the processing is necessary for reasons of "substantial public interest" could also be the basis for lawful processing of sensitive personal data, but since art 9.2(j) specifically refers to scientific research purposes, processing that takes place for scientific purposes should be based on the legal ground of art 9.2(j).
[46] Worth mentioning here is that in many countries, such processing by a university, even if consent is given by the data subject, could take place only after an ethics committee permits it. See also SOU 2017:50 s160.

animals, or even deceased persons (even though in this last case, Member States may provide for specific rules[47]).

3.2.1 Primary Versus Secondary Data Collection and the Principle of Transparency. An important conceptual and legal distinction resides in the selection of methods for data collection. For example, there is a significant difference between data collected directly from the data subject (e.g., small/medium-scale data obtained through surveys) and data collected through a third actor (e.g., online social networks obtained from APIs) without the direct involvement of the data subject. The difference here is not only in the scale or the nature of the data but in the relation between the data subject and the data controller: two different articles are concerned with providing information to the data subject when the data are collected directly from them[48] and when data about them have not been obtained from them.[49]

In essence, these articles detail some of the ways that the principle of transparency must be put into action. Transparency addresses the right of the data subject to know and understand how the data are being used; it "requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and [in] clear and plain language [in particular] in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected [. . .]." If personal data are collected, the data subjects should be informed about the collection and its purposes to enable them to exercise their rights. Note that this is different from consent (explained in Section 3.1) but instead refers to the information that must be made available about data processing activities. Essentially, data subjects should be able to easily find out who might be using their data and for what purposes.

Although making the data subjects aware of the processing and of their rights may seem straightforward when data are collected directly from them, this can become very difficult to accomplish when large networks are obtained from APIs. The potential difficulties to provide information under specific circumstances are acknowledged in the GDPR, where exceptions for research in particular are introduced. Article 14 states that providing information is not necessary if (1) "the data subject already has the information" or (2) "the provision of such information proves impossible or would involve a disproportionate effort, in particular for [. . .] scientific or historical research purposes," subject to some safeguards,[50] if providing information "is likely to render impossible or seriously impair the achievement of the objectives of that processing." Article 14 then continues stating that "[i]n such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available."

These are some examples of the kinds of research exemptions embedded in the GDPR, codifying and specifying research conduct. Both of those exemptions apply to social network research based on online data collected from social media platforms assuming that social media platforms have already informed their users through appropriate Terms of Services that their data will be shared with third parties (e.g., through APIs) or assuming that the large scale of collected data will require a disproportionate effort to inform all affected data subjects. This is an example of balancing research needs against the derogation of the rights of the data subject. Technically termed *proportionality of the effort*, this is a relatively vague concept. The controller, to determine whether

[47] Rec 27 GDPR.
[48] Art 13 GDPR.
[49] Art 14 GDPR.
[50] Art 89 GDPR.

it is going to be disproportionately difficult to provide the information, must take into consideration the number of data subjects, the age of the data, and if there are any appropriate safeguards already adopted.[51] If, after this assessment, the controller finds that the effort will be disproportionate, then she has to assess once again whether the effort involved to provide the information to the data subject exceeds the impact and effects on the data subject in the case where the information is not provided. This assessment has to be documented, and depending on the outcome, the controller may have to take extra measures (e.g., pseudonymization or anonymization if possible and appropriate).

As an example, this means that although the research exceptions may not technically require that every single Twitter user of the millions involved in any large-scale Twitter network research be notified that their data are used for research, the logic involved in deciding to collect data and skip the notification must be formally documented. This documentation must also demonstrate that appropriate storage, security, and pseudonymization techniques have been considered. In addition, it is unclear whether providing information to these users should be considered an impossible or very difficult task. In any case, the disproportionate effort that it would require to provide information to the data subjects shall be demonstrated by the data controller and is not something that should just be taken for granted.

The concept of transparency is particularly relevant in the context of social network research, as previously highlighted, for example, by Borgatti and Molina [4], and as such it requires a more extensive discussion. In particular, some additional details should provide a better description of the obligations of the data controller with regard to the provision of information. Three points are important here.

First, the data controller must always provide information at the time the data are obtained (if obtained directly from the data subject) or within a reasonable period after the data are obtained and no later than a month (if the data are obtained indirectly), as long as this is possible given the appropriate adherence to the research exemptions detailed earlier.

Second, if the controller changes the purpose of the processing, she must provide the information to the data subject prior to this processing.[52] For example, research data may have been collected for one purpose but the research question has shifted in the course of the data analysis, and these data will now be used for a different purpose. This then speaks to how precisely the information about processing must be specified. Looking at rec 33, even though referring to consent, we can conclude that the specification in case of research can be a bit more general (e.g., the general research area or part of the project, not specific analytical task). Therefore, changing data analysis approaches and even research questions may not require informing the data subject anew.

Related to the preceding information is the fact that if the change leads to further processing that is incompatible to the initial purposes, mere information of the change does not "whitewash" other obligations of the controller. According to art 5.1(b) GDPR, processing should comply to the purpose limitation principle. That means that as soon as the new processing is incompatible to the initial processing, the controller should either avoid the new processing or find a new lawful basis for it. There is, however, an exception with regard to research purposes, as in such case the further processing for such a purpose is considered to be compatible to the initial purpose.

[51] Rec 62.

[52] Rec 61. See also Opinion, where it is stated that in case the change is related to an incompatible further processing, informing about the change does not "whitewash" other obligations of the controller, such as finding another lawful basis for the changed processing or asking for new consent.

Third, the general principle does not assume that the methods and the analysis are known in detail at the moment of the data collection. However, the common practice in many areas of research where data are often collected with no specific hypothesis/evaluation framework becomes problematic because at least a limited explanation for the purposes of data processing is always necessary. The GDPR recognizes that it is not always possible to know from the beginning the entire scope of the research until the data are collected and used. Rec 33 (in case of consent) states that data subjects should be able to "consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose." Thus, some specification of the intended purpose is necessary, limiting but not entirely eradicating exploratory forms of data collection.

3.2.2 The Depth of Online Social Network Data and the Principle of Data Minimization. Whereas some network data can be collected directly in the form of network information (i.e., nodes and edges), many network datasets are obtained through processing of other types of data. For example, this is often the case in research based on social media such as Twitter. Network studies of Twitter can be based on the user-articulated following/followers structure, which can be considered direct network information. At the same time, we can build networks mapping communication processes, either explicit (replies, mentions) or implicitly specified such as by the usage of common hashtags [15]. To build this second type of network, researchers collect the content of users' posts and then extract and infer relational information. The problem arises if we consider the implications of collecting the content of the posts to build the network. Depending on the topic of posts, the type of content that is likely collected may vary but could include data revealing information that is not only identifying of natural persons but also includes sensitive data such as political affiliation and religious belief.

The GDPR makes a distinction between different types of personal data, such as data with regard to ethnicity and sexual preferences (the so-called sensitive personal data[53]), and for the processing to be considered lawful, the controller must respect the essence of data protection rights and follow suitable safeguards.[54] Notice that data that in combination with other data can lead to revealing sensitive data may also be considered as sensitive data. For example, a name in combination with a phone number, where each piece of data is not sensitive, may constitute sensitive data together if they probably reveal the ethnicity of a person. It is easy to see how the average stream of messages written by an average user might easily contain sensitive personal data or data that can be combined to reveal sensitive personal data about the data subject. Further, such data can be derived about persons simply from information produced by their connections. For example, it may be possible to ascertain a person's political affiliation if the majority of his connections explicitly communicate theirs.

Handling sensitive data is not forbidden, but before starting the data collection, researchers need to plan some safeguards. Under the GDPR, controllers may not process sensitive personal data except if the subject has provided her "explicit consent"[55] or the data "was manifestly made public by the data subject,"[56] or in case of research purposes.[57] Although one may consider using the concept of "manifestly made public" for special cases such as online social networks, where the information is publicly posted online by the users, we advise against this interpretation. In fact,

[53]In the context of GDPR, sensitive personal data is defined as "personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms."
[54]Art 9 GDPR.
[55]Art 9(2)(a) GDPR.
[56]Art 9(2)(e) GDPR.
[57]Art 9(2)(j) GDPR.

in the context of social media, as a consolidated body of literature has made clear, assuming when something is "manifestly public" is problematic [7] and a potentially serious breach of standard ethical research practices. On the contrary, the exemption in case of research purposes can be used, although only if processing is necessary, in accordance to art 89(1), based on Union or Member State law that shall be "proportionate to the aim pursued, respect the essence of the data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject." Moreover, it seems that profiling on the basis of personal data is forbidden unless there are "suitable safeguards."[58] For example, in Sweden, it was recommended that one such security measure can be considered the decisions of the relevant ethics committee.[59]

Finally, even if the data are not sensitive, the data minimization principle should still apply. Using again Twitter data as an example, when researchers collect information based on a hashtag, they can fetch data using the hashtag with another meaning, and so not related to the study, or data using the hashtag as was intended, but still including additional unwanted information. This means that researchers must put in place mechanisms that will effectively strip out unwanted data and delete it as soon as possible, acknowledging and documenting the process.

3.3 Data Analysis and Profiling

Social network analysis includes a wide range of data analysis tasks. Sometimes whole-network statistics are important, for example, to correlate the communication/interaction structure of a team or organization to its performance. Sometimes meso-level structures are of interest, for example, if we want to identify communities [5, 12, 13] or other relevant sub-structures such as online conversations [19, 29] inside a larger network. The identified groups can then also be used to classify individual actors, for example, assigning them to a given community or role. Other types of micro-level analysis involve the characterization of single actors, for example, when the most central or prestigious actors are identified [30]. When individuals are the object of the analysis, which is the case for most of the tasks listed earlier, an important concept to be considered is profiling.

The GDPR puts a special emphasis on the concept of profiling by specifying the definition and codifying acceptable practices. Accordingly, in the GDPR, profiling is composed of three main stages "a) collection of personal data; b) automated analysis to identify correlations; c) applying the correlation [the result of b)] to an individual to identify characteristics of present or future behaviour."[60]

Note that the notion of "automated analysis" is used in the GDPR in opposition to "manual." Although both types of processing are under the purview of the GDPR, profiling is necessarily automated. However, automated here would mean both the use of a statistical software for conducting any form of data analysis and the use of more complex approaches such as machine learning algorithms. Thus, any data analysis that includes computational assistance from software falls under automated analysis and thus can be classified as forms of profiling.

Given the preceding information, many (but not all) social network analysis tasks can be classified as profiling. All centrality measures are clear examples, as they associate results of the network analysis to specific individuals. Any analysis that singles out individuals based on the identification of positions, roles, and communities is similarly a form of profiling.

What is the researcher to do if his activities constitute profiling of the data subject? This does not mean that the particular data analysis is disallowed. However, this may require the performance of

[58] Rec 51 GDPR.

[59] SOU 2017:50.

[60] Art 29 Data Protection Working Party, WP251rev.01, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679."

a data protection impact assessment (DPIA), for which the advice of the appointed data protection officer should be sought. Although the GDPR states that profiling has to be systematic and extensive to require a DPIA, many authorities have made a broader implementation and if profiling may affect individuals in general (e.g., it provides custom access to services, it includes sensitive data, it is related to vulnerable individuals, and in general the processing can lead to a high risk to the rights and freedoms of the data subject), and if it is conducted in a large scale combining sensitive data, then a DPIA is in general necessary. The question of whether a DPIA is necessary is clearly a very important one, because a very strict approach leading to an assessment for every possible case of social network analysis can become practically problematic for the researchers. While we wait for more guidelines[61] and other legal specifications, the role of the researchers together with the DPOs deciding on whether an assessment is needed or not (following the law but also being practical) is of even higher importance.

Alongside profiling, DPIAs are also applicable to systematic monitoring of individuals and locations. An interesting question arises with respect to what constitutes locations and public spaces. For example, the GDPR mentions a "systematic monitoring of a publicly accessible area on a large scale" as a reason for a DPIA.[62] We are not aware of existing legal interpretations of whether, for example, Twitter is a publicly accessible area, but the WP29 interprets "publicly accessible area" as being any place open to any member of the public, such as a piazza, a shopping center, a street, or a public library. Clearly, these are examples of physical places, but Twitter is also a place that is open to any member of the public provided that she has the means to access it (an Internet connection and access to an email address). Such questions will likely be decided later on as the regulation stands the test of time and litigation, but it is an important item to consider for researchers conducting large-scale collection and processing of ostensibly "public" data.

3.4 Data Storage

In this section, we discuss what happens after the research is concluded, in case the researchers want to store the collected networks. If the data are still personal, for example, they still contain identifiers or have been pseudonymized, then the data controller must guarantee some rights to the data subjects if she wants to keep the network data. On a general level, we can organize these rights along three lines: (1) temporal duration of personal data storage, (2) the accessibility of the stored data to the data subject, and (3) the right of the data subject to withdraw his or her data. All of these tasks are in general strictly regulated by the GDPR but with significant exemptions for research, discussed in the following section. Under the assumption that the networks have been anonymized, then there is no problem because the GDPR no longer applies: the data are no longer personal. However, network anonymization is a complex issue, which we also discuss in the following.

3.4.1 Rights of the Data Subjects and the Principle of Storage Limitation. When it comes to temporal storage limitation, the GDPR states that in general data can be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed," but more extended periods may apply in case of research as long as appropriate safeguards are followed.[63]

No exemption because of research is instead mentioned regarding the data subjects' right to check if there are data concerning them and the right to obtain these data.[64] This means that

[61] https://www.ucl.ac.uk/legal-services/research/data-protection-impact-assessment.
[62] Art 35(3)(c) GDPR.
[63] Art 5(e) GDPR.
[64] Art 15 GDPR.

when requested, the controller should provide the data in a "commonly used and machine-readable format"[65] (even if there are possibly other national laws that may restrict this right of a data subject, such as secrecy acts[66]). Considering the average amount of data represented by a single node in a typical social network project, this should not be a problem. Nevertheless, as for other parts of this article, the size of the network may constitute a practical difference, and for large networks researchers should probably consider implementing an automated data filtering functionality.

Finally, the right to erasure, also known as the right to be forgotten, grants to the data subject "the right to obtain from the controller the erasure of personal data concerning him or her without undue delay."[67] However, also in this case, the GDPR contains an exemption to this obligation if the erasure "is likely to render impossible or seriously impair the achievements of the objectives of that processing."[68] Many SNA measures are not so sensitive to a small amount of missing data [16], and the discipline has developed a set of techniques to handle missing data. Nevertheless, it should be acknowledged that a significant number of subjects requesting their data to be removed might seriously impair the research objectives, and thus researchers would have the right to legally object to the data removal.

Although these are the general guidelines emerging from the GDPR, according to art 89(2), Member States may further limit the data subjects right to access, rectification, and restriction, and to object in case of research if there are appropriate safeguards in place, and as long as the derogation is necessary for the fulfillment of the research.

3.4.2 Data Anonymization. The GDPR asks for appropriate safeguards. The safeguards that are named in the GDPR are technical and organizational, such as data minimization, pseudonymization, and anonymization. In addition, there can also be legal safeguards, such as contractual clauses between the controller and the processor and ethical vetting [20]. Here, we focus on anonymization, which should result in the data not being re-identifiable by the controller or any other person. In social network analysis, the typical approaches to anonymization are based on clustering, graph modification, or network perturbation [31].

Data anonymization approaches in general are part of a considerable debate where some researchers argue that anonymization is impossible, whereas others contend that it is only in some cases [1, 10, 23]. Data from social networks are far more difficult to anonymize than other types of data, and research on appropriate anonymization techniques is still in its relative infancy. Many of the simpler and more traditional approaches, such as replacing node identifiers, and more recent and complex approaches have been critiqued as insufficient [1, 2]. The knowledge of research being conducted in a particular location by a specific research group may be enough to reveal the identities of individuals encoded in the network to those who are familiar with these people more directly. In a small social network, such as a company division, it may be simple for the people in the network to recognize others based on just the revealed relational patterns [4]. Such an issue is not specific of social networks but has been amply documented in qualitative and ethnographic research [27, 28]. As another case, if the data are public and indexed (e.g., by Web search engines), it can be very easy to find the original data using a part of it as a search key, such as finding the authors of a social media post based on the text of the post.

Whether anonymization or even just pseudonymization are generally possible in a social network context is a difficult question. The GDPR states the necessity for privacy by design and by default but does not request specific privacy-preserving solutions: the controller should select and

[65] Art 20 GDPR.
[66] SOU p223.
[67] Art 17 GDPR.
[68] Art 17 GDPR.

apply the appropriate measures for each case. In the GDPR, pseudonymization requires the "additional information" to be "kept separately" and to be "subject to technical and organisational measures,"[69] which is not really possible when the data source is public: if one removes the user identifier but keeps the text of the post (e.g., the tweet), a simple search on a search engine or on the social media platform can easily lead to the original, complete information. In this case, a possibility to be considered by the researchers (but not explicitly required by the GDPR) is to transform the text so that the analysis can still be performed, but it becomes more complicated to fetch it from the Web, such as replacing it with a bag of words. The relevance of this discussion is that according to rec 26, pseudonymized data are identifiable, so the GDPR applies to that, whereas anonymized data is not, so the GDPR is no longer relevant. However, given the difficulty in fully anonymizing the data, we should often assume that the GDPR is still the relevant regulation.

Even when we do not need identifiers to process social network data, because, for example, we are only interested in the structure of the network and its relationship with some indicators, we still need the identifiers if we want to extend the network to know to which nodes the newly available information refers. According to the GDPR, we should at the very least pseudonymize the data "as soon as possible" (recital 78). However, it is not unusual in online network studies to keep collecting data for months or even years, which means that "as soon as possible" may be as late as the end of the study. One solution here is to develop or extend data collection systems with built-in network pseudonymization functions, such as automatically removing identifiers and separately storing a mapping to user accounts in a location that requires special access credentials. Such solutions may seem overly onerous given the current accepted practices, but the GDPR forces us to rethink our attitudes toward data collection and the impact of our practices more broadly. In addition, the idea of designing ethically related features in social network analysis software has already appeared in the literature [4].

As a final note, whereas in previous paragraphs we discussed the difficulty of network data anonymization, there are specific types of social network data where anonymization is indeed possible. In ego-network data collection, different actors are asked about their own social ties and perhaps those of their neighbors. Ego-networks are then analyzed without reconstructing a common network for all of the participants. In this case, there is typically no need to know who the individuals are, which means that we can design a data collection that is already anonymized at the source. As a result, these data are outside the scope of the GDPR given the definition of anonymous data as "data rendered anonymous in such a way that the data subject is not or no longer identifiable."[70]

4 FLOWS OF DATA

So far we have considered cases where only one data controller processes the data, as in the case of a single research team based at one institution performing the research. In practical situations it can however happen that data is stored by a team at a university and sent to a team at another university to be analyzed, or that two universities perform a joint data collection.

It is not the goal of this article to examine the legal implications regarding dataflows from one jurisdiction to another, but there are some things that are worth naming here. First, when it comes to transfers of personal data within the EU, the GDPR has as a goal to "prevent divergences hampering the free movement of personal data within the internal market."[71] However, when it comes

[69] Art 4(5) GDPR.
[70] Rec 26 GDPR.
[71] Rec 13 GDPR.

to data related to research, the situation is somewhat more complicated because quite many issues are left to the Member States to decide.[72]

Moreover, regarding transfers to third, non-EU, countries, it has to be made sure that such transfers comply to the safeguards provided for in art 44 et seq GDPR. We will not analyze the different possibilities for such compliance to be achieved, but it is important to emphasize that when it comes to transfers of data to entities in third countries, the situation is far from problem free. By way of illustration, such transfers are allowed if the data are sent to a "safe country," namely to a country recognized as a country providing an equally adequate level of data protection as the EU countries. However, for the time being, these countries are limited to a handful of—mostly minor—countries.[73] Alternatively, such transfers must be based on the consent of the data subject, something which, however, as already stated earlier, can be difficult to obtain in cases of research in networks. Similarly, transfers to third countries are allowed if the data-transferring party and the data-receiving party use an EU Standard Contractual Clause. However, such clauses have already been challenged with regard to their ability to provide an adequate level of protection of personal data.[74]

5 SOME MORE GENERAL ISSUES AND CONSIDERATIONS

5.1 A Code of Conduct for Social Network Research

The opportunity of writing a code of conduct for research in social network analysis has been under discussion for a long time. In the special issue of *Social Networks* on ethical dilemmas in social network research, there was mention to "efforts now underway within INSNA, the professional association for social network researchers, to establish a code of ethics" [8]. Several years later, at a board meeting of the same association,[75] this was still under discussion, and it was noted that many members of the association are also members of other associations for which codes of conduct already exist (e.g., by the American Anthropologist Association, the American Political Science Association, the American Sociological Association, and the Association for Internet Researchers (AOIR)), questioning the need for an additional effort.

We believe that this article can contribute to this discussion. On the one hand, we note that many issues highlighted in the previous sections are common to other types of research not necessarily involving social networks, including, for example, social science research in general, Internet research, and big data analysis, even though some specific aspects of social networks have also emerged and the combination of relevant issues is also unique. On the other hand, the broad picture emerging from our analysis of the GDPR is a complex one, and a whole section of the regulation[76] indicates codes of conducts as a way to reduce this complexity. In fact, once a code of conduct proposed by an association has been approved, registered, and published by a supervisory authority certifying that the code is compliant with the GDPR and "that it provides sufficient

[72] Art 89 GDPR. For a short analysis on the matter, see also C. Staunton, S. Slokenberga, and D. Mascalzoni, The GDPR and the Research Exemption: Considerations on the Necessary Safeguards for Research Biobanks *(European Journal of Human Genetics* 2019), available at http://www.nature.com/articles/s41431-019-0386-5.

[73] Andorra, Argentina, Canada (commercial organizations), Faeroe Islands, Guernsey, Israel, Isle of Man, Jersey, New Zealand, Switzerland, Uruguay, and the United States but only with respect to the Privacy Shield–certified companies. See https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.

[74] See reference for a preliminary ruling from the High Court (Ireland) made on May 9, 2018—*Data Protection Commissioner v. Facebook Ireland Limited, Maximillian Schrems* (Case C-311/18).

[75] 2016 Report to INSNA Membership prepared by the INSNA officer.

[76] Sec 5, art 40-43.

appropriate safeguards,"[77] then showing compliance with the code exempts the data controller from a number of obligations. In summary, since the enforcement of the GDPR, the benefits of codes of conducts have increased, but their establishment requires additional effort because they require an authority to verify their compliance.

5.2 Toward GDPR-Compliant Social Network Software

Through the analysis of the legal obligations emerging from the GDPR, we have seen many cases where the law can be considered a bottom line for ethics, where individual researchers shall consider more restrictive actions. For example, as we discussed earlier, the GDPR explicitly mentions "disproportionate effort" as a reason not to provide information to the data subjects. This, when framed within the context of online data or of secondary analysis of already collected large datasets, might easily be used as a solid reason to perform research without informing the data subjects. But if it is true that large online data could easily count millions of potential data subjects, one can expect that for online sources it can be possible to automatically send notifications or messages informing the data subjects. Although this may result in a potentially significant overhead of communicating with confused data subjects, the effort may be a first step in acknowledging that people who produce data must be treated with dignity and respect regardless of research aims. Development of standards for notification in large-scale data collection endeavors is necessary and may need to be taken up at the level of professional codes of conduct.

In these cases, we should also consider developing tools that can take care of notification automatically, reducing the claim of disproportionate effort rather than leveraging it as a way to sidestep responsibilities in research. For example, in the growing context of Twitter research, sending a short tweet mentioning those user accounts included in the social network data collected in a research project would be potentially interesting information for data subjects, contributing to the creation of an awareness about how much our public data is used. If done by a relevant share of researchers (which theoretically can be achieved if the main tool or tools for data collection are extended with this functionality), this increased awareness could result in a consequential generalized improvement in the way people manage their data online and an increased trust in science, showing how careful researchers are about this. However, although automatically sending the information to a list of users seems to require a limited effort, turning this into practice can be problematic, as described in the next section.

5.3 An Experiment on Automated Data Subject Information on Twitter

To better understand the amount of effort needed to notify data subjects in the context of online social network research, we have set up a protocol for a Twitter data collection process. This experiment, briefly reported in the following, highlighted the difficulty of performing even a task (appearing to be) as easy as sending some information to online users.

First, we had to consider a number of alternatives. First, when tweets are collected on Twitter, the only contact information we have are the Twitter identifier and screen name of the accounts whose tweets were collected. This means that we can only inform Twitter users via Twitter, and as it is typically not possible to send direct (private) messages to generic Twitter users (e.g., users not following us who have not explicitly allowed this in their privacy settings), we need to inform them in some way that is visible to others, such as using a public mention.

Q2

Although this is not necessarily problematic, it is interesting to see how to inform data subjects so that we can protect their privacy we have to release additional information about them: our public message implies that those accounts have posted tweets with the hashtag we were

[77] Art 40(5).

monitoring, that by the principle of transparency we have to clearly indicate in the communication. Then, we must decide (1) how many users to mention in the same tweet and (2) whether we should check what their current screen name is. Both choices have an impact on the time needed to send the notifications: including more accounts in the same Tweet would reduce notification time but would also again release more information as each notified user would see the other user names in the same tweet, knowing that they have also used the same hashtag. Checking the current screen name would require more accesses to the Twitter API but would avoid that we mention the wrong account because screen names can change in time. As an indication, the Twitter API currently allows us to send 2,400 tweets per day, meaning that we would need around 1 year and 2 months to notify 1 million users (using a single notification account).

An alternative is to notify users through the hashtag, sending a tweet without mentions but with the monitored hashtag and specifying that we are collecting tweets containing the same hashtag. This, however, is also problematic, first because there is no guarantee that users will see that (they would have to search tweets containing that hashtag, at the right time), and second because for some studies awareness of the data collection might result in a different behavior.

Other decisions making the practical information process less trivial than one may think are whether we should also notify accounts mentioned in the collected tweets, even if they were not producing tweets themselves, or whether we should notify accounts retweeting other accounts' tweets.

After deciding on all of these aspects, we started sending our tweets, including a link to the information about the project and the data processing, as well as information about the user rights. The procedure for the users to offer them (among other things) the possibility of retrieving the data about them we had collected was also complicated, because to prove their identity the users were requested to follow our notification account, which is again revealing more information about the user and also requires some effort that might discourage potentially interested users. After sending notifications to 45 accounts, we registered only one visit to the information page.

Finally, Twitter blocked our notification account. According to their rules, the account had been marked as having a spamming behavior. In the process to reactivate the account, we mentioned that despite the behavior being compatible with their definition of spamming, the account was an attempt to enforce the rights of the users to know that their tweets had been collected and why, but this did not result in any exception, which led us to drop the experiment after considering that developing a "smarter" bot sending the notifications and trying to behave in a way not to be caught by Twitter's algorithms would have been ethically questionable.

5.4 Public Versus Private Actors

Another important point of discussion is the difference in classification of universities and commercial research and industry labs. The stark difference in legal basis for data processing and the impact on the consideration of whether consent is a legitimate lawful basis is an important point to consider. What does consent constitute in the context of a commercial entity when it must be clearly uncoerced and freely given? What are the different obligations toward data subjects for researchers depending on the legal ground they employ for data collection? These questions have some answers in the GDPR but will be further evolving as time and litigation test the GDPR terms and definitions. Data availability for large-scale computational social science and social network research is necessarily connected to commercial actors [18]. Collaborators across academic and commercial spheres have claimed the unalloyed public good that is possible from large-scale data collection, but what impact may the GDPR have, given the differentiation it makes between public and commercial research efforts? How much access will public university researchers continue to

have to commercial data stores? How complex will these negotiations become? These questions are beyond the purview of this article but must be discussed and considered in the future.

6 CONCLUSION

Our main objective when we started writing this article was to provide a practical guide to GDPR-compliant social network data processing. Working on it, and also trying to apply our recommendations to our own research, it became evident that although some issues could be more easily translated into practical suggestions, other general indications and principles in the regulation are difficult to either interpret or apply in the context of social network research. The problems we have highlighted in the article include the difficulty of sending information to millions of users through a third-party API that does not allow it, the problems in pseudonymizing the data as soon as possible in a continuous network monitoring process performed with pre-GDPR software tools, the interpretation of concepts such as "manifestly made public data" and "publicly accessible areas," the problem of removing data by user request not knowing what impact this will have on network statistics, the practical impossibility of guaranteeing respondent anonymity, the inclusion of data about individuals not included in the study, and more general issues related to data protection emerging when social network analysis is applied to large-scale networks of social relations derived from social media data.

In summary, it is important that everyone involved in the processing of social network data invests some time to reflect about the implications of the GDPR on their research, seeking help from their institutions but not only relying on institutional support. Although this may sound like an obvious statement, and legal and ethical problems related to social network analysis and Internet research for social behavior have certainly received a lot of attention in the past as witnessed by the literature on the topic and by existing codes of conduct, the sudden explosion of online behavioral data has indeed affected the research landscape by both introducing new problems and involving new researchers from disciplines where some of these problems had not been traditionally accounted for.

REFERENCES

[1] A. Narayanan and E. W. Felten. 2014. *No Silver Bullet: De-identification Still Doesn't Work*. Technical Report. Princeton University.

[2] L. Backstrom, C. Dwork, and J. Kleinberg. 2007. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*. 181–190.

[3] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012), 295.

[4] S. P. Borgatti and J.-L. Molina. 2005. Towards ethical guidelines for ethical research in organizations. *Social Networks* 27, 2 (2005), 107–117.

[5] Cecile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova. 2015. Clustering attributed graphs: Models, measures and methods. *Network Science* 3, 3 (2015), 408–444.

[6] D. Boyd. 2008. *Taken Out of Context: American Teen Sociality in Networked Publics*. Ph.D. Dissertation. University of California-Berkeley, School of Information.

[7] D. Boyd. 2010. *Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. A Networked Self*. Routledge.

[8] R. L. Breiger. 2005. Introduction to special issue: Ethical dilemmas in social network research. *Social Networks* 27, 2 (2005), 89–93.

[9] A. M. Capron. 2018. Where did informed consent for research come from? *Journal of Law, Medicine & Ethics* 46, 1 (2018), 12–29.

[10] A. Cavoukian and K. El Emam. 2011. *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy*. Technical Report. Information and Privacy Commissioner of Ontario, Canada.

[11] G. Chassang. 2017. The impact of the EU general data protection regulation on scientific research. *Ecancer* 11 (2017), 709.
[12] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. 2011. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining* 4, 5 (2011), 512–546.
[13] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3–5 (2010), 75–174.
[14] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779.
[15] Obaida Hanteer, Luca Rossi, Davide Vega D'Aurelio, and Matteo Magnani. 2018. From interaction to participation: The role of the imagined audience in social media community detection and an application to political communication on Twitter. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'18)*. IEEE, Los Alamitos, CA, 531–534.
[16] Gueorgi Kossinets. 2006. Effects of missing data in social networks. *Social Networks* 28, 3 (2006), 247–268.
[17] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. Computational social science. *Science* 323, 5915 (2009), 721–723.
[18] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, and T. Jebara. 2009. Life in the network: The coming age of computational social science. *Science* 323, 5915 (2009), 721–723.
[19] Matteo Magnani, Danilo Montesi, and Luca Rossi. 2012. Conversation retrieval for microblogging sites. *Information Retrieval* 15, 3–4 (2012), 354–372.
[20] C. Magnusson Sjöberg. 2017. Scientific research and academic e-learning in light of the EU's legal framework for data protection. In *New Technology, Big Data and the Law (Perspectives in Law, Business and Innovation)*, M. Corrales, M. Fenwick, and N. Forgo (Eds.). Springer, Singapore, 43–65.
[21] A. Mantelero. 2014. The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics. *Computer Law & Security Review* 30, 6 (2014), 643–660.
[22] L. Marelli and G. Testa. 2018. Scrutinizing the EU General Data Protection Regulation. *Science* 360, 6388 (2018), 496–498.
[23] A. Narayanan, J. Huey, and E. W. Felten. 2016. A precautionary approach to big data privacy. In *Data Protection on the Move*, S. Gutwirth, R. Leenes, and P. De Hert (Eds.). Springer, Heidelberg, Germany, 357–385.
[24] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY.
[25] S. Penasa, I. de M. Beriain, C. Barbosa, A. Białek, T. Chortara, A. D. Pereira, P. N. Jiménez, T. Sroka, and M. Tomasi. 2018. The EU General Data Protection Regulation: How will it impact the regulation of research biobanks? Setting the legal frame in the mediterranean and eastern european area. *Medical Law International*.
[26] K. Schaar. 2016. What is important for data protection in science in the future? General and specific changes in data protection for scientific use resulting from the EU General Data Protection Regulation. In *Working Paper Series of the German Council for Social and Economic Data*. German Council for Social and Economic Data (RatSWD).
[27] I. Shklovski and J. Vertesi. 2013. "Un-Googling publications": The ethics and problems of anonymization. In *Proceedings of CHI'13 Extended Abstracts on Human Factors in Computing Systems(CHI EA'13)*. 2169–2178.
[28] W. C. van den Hoonaard. 2003. Is anonymity an artifact in ethnographic research? *Journal of Academic Ethics* 1, 2 (2003), 141–151.
[29] Davide Vega and Matteo Magnani. 2018. Foundations of temporal text networks. *Applied Network Science* 3, 1 (2018), Article 25, 26 pages.
[30] Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.
[31] Jin Zhou, Xiaoke Xu, Jie Zhang, Junfeng Sun, Michael Small, and Jun-An Lu. 2008. Generating an assortative network with a given degree distribution. *International Journal of Bifurcation and Chaos* 18, 11 (2008), 3495–3502.

Received March 2019; revised June 2019; accepted October 2019

Author Queries

Q1: AU: Please supply complete mailing addresses for the authors.

Q2: AU: Please confirm that sentence is clear: "Although this is not necessarily . . . "

Q3: AU: Please provide a volume, issue number, and page range for reference 25. Unable to locate.