Data Confidentiality and Information Credibility in Online Ecosystems

Giovanni Livraga
Università degli Studi di Milano
Dipartimento di Informatica (DI)
Via Giovanni Celoria, 18
Milan, Italy
giovanni.livraga@unimi.it

Marco Viviani
Università degli Studi di Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo) – Viale Sarca, 336
Milan, Italy
marco.viviani@unimib.it

ABSTRACT

Recent ICTs paradigms such as cloud computing, data outsourcing, digital data markets, and the spread of multiple social media based on Web 2.0 technologies, facilitate the exchange of large data and information flows among a myriad of interconnected devices and users, for different aims and purposes. This complex scenario underlies the development of online ecosystems of interacting entities, where the concepts of community, self-organization, evolution and knowledge are fundamental.

While the benefits connected to such kind of ecosystems are intuitive also to the everyday man, no lunch comes for free, and such a complex and interconnected scenario entails a number of issues connected to both data and information generation and diffusion that should be carefully addressed. For example, in the data sharing context, genuine data could be manipulated, tampered with, accessed without permission, breached, or improperly disclosed; in the Social Web context, low-quality data and/or misinformation could be diffused. With respect to the above-mentioned issues, in this paper we survey some of the possible approaches proposed in the literature for ensuring adequate data protection, with particular reference to data confidentiality, and for assessing information credibility in complex online environments. We also provide a conclusive discussion aimed at illustrating the importance of relating these concepts.

CCS CONCEPTS

• **Information systems → World Wide Web**; • **Security and privacy → Pseudonymity, anonymity and untraceability**; *Social aspects of security and privacy*; Database and storage security.

KEYWORDS

Digital Ecosystem; Data Sharing; Data Protection; Confidentiality; Social Media; Credibility

MEDES '19, November 12–14, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-6238-2/19/11…$15.00
https://doi.org/10.1145/3297662.3365829

ACM Reference Format:
Giovanni Livraga and Marco Viviani. 2019. Data Confidentiality and Information Credibility in Online Ecosystems. In *11th International Conference on Management of Digital EcoSystems (MEDES '19), November 12–14, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297662.3365829

1 INTRODUCTION

A *Digital Ecosystem* can be defined as a "distributed, adaptive, open socio-technical system with properties of self-organisation, scalability and sustainability inspired from natural ecosystems" [5]. Based on this definition, recent ICTs paradigms exploiting Web 2.0 technologies are at the basis of the development of interactive, hyper-connected, immersive, virtual, complex *online ecosystems* where users create, collect and share data, information, and knowledge for different purposes, often in a collaborative way [4]. For instance, cloud computing can facilitate storing, sharing, and analytics of different kinds of data, while the Social Web allows users to cooperate in virtual communities for building and sharing knowledge (e.g., Wikipedia), interacting with each other (e.g., LinkedIn, Facebook), and, in general, communicating globally by means of microblogging sites (e.g., Twitter), photo and video sharing sites (e.g., Instagram, YouTube), etc. [28].

Despite its intuitive benefits, this complex and interconnected scenario entails a number of issues that should be carefully addressed. Considering the data sharing context, assuming genuine data and data sources, the attention must be placed on the fact that personal and/or sensitive data might not be freely shared due to *confidentiality* reasons, as also demanded by recent laws and regulations such as the European General Data Protection Regulation (EU GDPR) [3]. Conversely, the social context is characterized by an increasing volume of *User-Generated Content* (UGC) diffused by possibly unknown and uncontrolled sources of information [6], where the *credibility* of the content must be assessed [34, 41, 47]. Confidentiality and credibility issues, while appearing at a first sight complementary and orthogonal, are in fact strictly connected, as they may represent obstacles in pursuing the goals of an online ecosystem, for example to take knowledge-based decisions operating on data and information analytics. Without effective protection approaches, data sharing would not be possible; without effective credibility assessment solutions, misinformation and consequent wrong decisions could be generated.

In this article, we discuss the above-mentioned issues, and we illustrate the main state-of-the-art privacy definitions and approaches for protecting data confidentiality in the data sharing

context (Section 2), and for assessing information credibility in the Social Web context (Section 3). We also discuss, at the end of the paper (Section 4), the open issues related to both the considered problems, and the benefits that a uniform approach blending both confidentiality and credibility could have on online ecosystems.

2 DATA PROTECTION

Sharing data is a key enabler for any digital (online) ecosystem. However, whenever a dataset is to be shared among different agents in the ecosystem, major issues related to data protection can arise [15]. Among these, ensuring adequate protection to personal/sensitive information. In this section, we discuss some issues and possible approaches related to the protection of data confidentiality. For the sake of readability, we will refer to structured datasets (relational tables defined over a set of attributes of interest) that contain personal/sensitive information related to individuals, called *data respondents*. However, the discussed problems and techniques hold also with other data formats and with any generic sensitive information. In the remainder of this section, we will refer our examples to a medical (online) ecosystem, where different agents (hospitals, universities, pharmaceutical companies, medical insurances, and their representatives) collect and share medical data about patients for advancing research.

2.1 Confidentiality in Data Sharing

Protecting sensitive and personal information in *data sharing* scenarios typically requires the dataset owner (i.e., the agent owning the dataset and wishing to share it) to apply some modifications to it to ensure that no personal/sensitive information be improperly disclosed to the recipients. The research community has devoted major efforts in the development of techniques and approaches for properly protecting the *confidentiality* of a dataset to be shared. Clearly, the protection approach to be used depends on the specific scenario and on the privacy requirement to be satisfied. The main *privacy definitions* that have been investigated by the research community can be broadly classified in two families, namely *syntactic* and *semantic definitions*, as follows [14].

- *Syntactic privacy definitions.* They capture the protection degree enjoyed by a dataset with a numerical value (e.g., each release of data must be indistinguishably related to no less than a certain number of individuals in the population). Protection approaches pursuing a syntactic privacy requirement have been proposed to protect the identities of data respondents, and to break the correspondence between a respondent and her sensitive information (or, more generally, sensitive associations among data that should be protected). These approaches rely on a precise identification of what is sensitive and what are the sensitive associations to be protected, and typically guarantee data truthfulness thanks to the application of non-perturbative data protection techniques.
- *Semantic privacy definitions.* They model a property to be satisfied by the mechanism chosen for releasing the data (e.g., the result of an analysis carried out on a released dataset must be insensitive to the insertion or deletion of a record in the dataset). Protection approaches pursuing a semantic privacy requirement aim at releasing information that is slightly distorted, so to hide the actual informative content, and, for this reason, do not guarantee data truthfulness (although they produce, to some degree, reasonable results). They can be applied to produce a sanitized version of a dataset to be protected, as well as to release sanitized answers to queries posed agains the original (unprotected) dataset.

SSN	Name	DoB	Sex	ZIP	Disease
		1950/09/10	F	97401	Stroke
		1960/03/12	M	98302	Flu
		1950/09/04	F	97467	Colitis
		1960/03/20	M	98245	Flu
		1960/07/12	M	98312	Gastritis
		1950/09/11	F	97434	Asthma
		1960/07/25	M	98223	Cancer
		1960/07/30	M	98389	Cough
		1960/03/12	M	98290	Flu
		1940/12/01	*M*	*97210*	*Lupus*

(a)

Name	Address	City	ZIP	DoB	Sex
…	…	…	…	…	…
John	1100 Main Street	Portland	*97210*	*40/12/01*	*male*
…	…	…	…	…	…

(b)

Figure 1: An example of de-identified (a) and of publicly available non de-identified (b) datasets.

Whatever the privacy requirement, ensuring adequate protection is far from being a trivial problem. For instance, the naive solution of de-identifying a dataset, that is, the removal of all identifying information (e.g., names, e-mail addresses, social security numbers) is unfortunately not sufficient to guarantee anonymity. In fact, a de-identified dataset can still include other information, called *quasi-identifier* (QI), which can be linked to external sources to reduce the uncertainty about the identity of some respondents [14]. To illustrate, consider the de-identified dataset in Figure 1(a), to be shared in our ecosystem. The excerpt of a non de-identified dataset (e.g., a voter list) in Figure 1(b) contains a single record of a male, born on 1940/12/01, living in 97210. If this combination of values is unique in the external world as well, then the two datasets can be linked, disclosing the fact that the last record of the de-identified dataset pertains to *John*, and also the fact that *John* suffers from lupus. Based on a study performed on the US 2000 Census, Golle discovered that 63% of the entire US population is *uniquely identifiable* by the combination of their gender, ZIP, and full date of birth [21].

Intuitively, removing also QI information besides the direct identifiers might not be a feasible solution, since QI could represent a large portion of the informative content of the dataset to be shared and hence its complete removal would make the dataset useless for recipients. In the remainder of this section, we illustrate some of the main approaches for ensuring confidentiality protection when a dataset is to be shared and/or released. We refer to protection

approaches pursuing a syntactic (semantic, respectively) privacy definition as syntactic (semantic, respectively) approaches.

2.2 Syntactic Approaches

Syntactic approaches typically aim at counteracting the re-identification attack illustrated in Section 2.1. They can be based on the adoption of data generalization and/or data fragmentation, and typically guarantee the truthfulness of the protected data, which is simply less precise or less complete. We now illustrate some of the main protection approaches based on generalization and on fragmentation.

2.2.1 Generalization-based. Data generalization is a non-perturbative technique by means of which a data value is replaced with another, more general, value. For instance, an individual's complete date of birth ⟨`year/month/day`⟩ can be generalized to ⟨`year/month`⟩, or just to ⟨`year`⟩. By reducing the level of details of the dataset to be shared, the probability of finding unique correspondences with external data sources (see example in the previous section) clearly diminishes, reducing the risk of re-identification.

The first approach in this direction is represented by k-anonymity [39]. k-Anonymity enforces a syntactic privacy requirement (typically adopted by statistical agencies) demanding that *each release of data should be related indistinguishably to no less than a certain number of individuals*. This requirement would, in principle, require knowledge of all possible external data sources that could be linked to the dataset to be released, which is clearly not a reasonable assumption. To solve this issue, k-anonymity takes a safe approach and demands that *each combination of QI should appear with 0 or at least k occurrences in a released dataset*. Intuitively, this suffices to the satisfaction of the indistinguishability requirement, since each individual in any external data source could be mapped to 0 or greater than k records in the anonymized dataset. To achieve this goal, k-anonymity removes direct identifiers, and generalizes the QI according to specific generalization strategies [39].

The dataset in Figure 2(a) represents a k-anonymous version of the dataset in Figure 1(a) with k=3. Note that the last record of the original dataset (i.e., the one related to John) has been suppressed (i.e., removed). Suppression is adopted by k-anonymity in conjunction with generalization to reduce the amount of generalization that would otherwise be needed [31]. Generalization and suppression can be adopted at different granularity levels, and their combination defines different approaches for enforcing the k-anonymity requirement. For instance, the dataset in Figure 2(a) is produced adopting suppression at the level of entire records, and generalization at the level of entire columns.

The original definition of k-anonymity has then been extended to counteract specific attacks to which a k-anonymous dataset could be vulnerable. For instance, consider the first equivalence class (i.e., the set of $\geq k$ records sharing the same QI values) in Figure 2(a). If a recipient knows that a target respondent is in that equivalence class, then she is also able to discover that the respondent suffers from *flu*, even without discovering which is the actual record. To counteract similar issues, the original definition of k-anonymity has been extended to more complex definitions, such as those of ℓ-diversity [32] and t-closeness [30], so to take into consideration

SSN	Name	DoB	Sex	ZIP	Disease
		1960/03/**	M	98***	Flu
		1960/03/**	M	98***	Flu
		1960/03/**	M	98***	Flu
		1950/09/**	F	97***	Stroke
		1950/09/**	F	97***	Colitis
		1950/09/**	F	97***	Asthma
		1960/07/**	M	98***	Gastritis
		1960/07/**	M	98***	Cancer
		1960/07/**	M	98***	Cough

(a)

SSN	Name	DoB	Sex	ZIP	Disease
		1960/**/**	M	983**	Flu
		1960/**/**	M	983**	Cough
		1960/**/**	M	983**	Gastritis
		1950/**/**	F	974**	Stroke
		1950/**/**	F	974**	Colitis
		1950/**/**	F	974**	Asthma
		1960/**/**	M	982**	Flu
		1960/**/**	M	982**	Flu
		1960/**/**	M	982**	Cancer

(b)

Figure 2: An example of 3-anonymous (a) and 2-diverse (b) versions of the dataset in Figure 1(a).

the sensitive values when clustering records in the equivalence classes for ensuring the k-anonymity requirement. For instance, ℓ-diversity requires each equivalence class to contain at least ℓ well-represented values for the sensitive attribute, so to counteract the issue mentioned above. The dataset in Figure 2(b) represents an example of a 2-diverse (and 3-anonymous) version of the dataset in Figure 1(a), where each equivalence class counts at least two different values for the sensitive attribute.

2.2.2 Fragmentation-based. While generalization indeed guarantees data truthfulness, it produces incomplete/imprecise QI information. In some scenarios this might be problematic, for instance when the QI need to be precisely analyzed. An alternative strategy to enforce the protection guarantees illustrated above is to adopt data fragmentation. Fragmentation consists in splitting the original dataset in a set of fragments (i.e., vertical views over the original dataset) to break the correspondence between data that should not be visible together, such as the QI and the sensitive data, and in releasing information on the association at the level of groups of records to create confusion on the real original associations. For instance, Anatomy [48] is a fragmentation-based proposal that achieves ℓ-diversity without resorting to generalization. This approach first partitions the records in the dataset in groups such that each group contains at least ℓ well-represented sensitive values, and assigns an identifier to each group. The dataset is then split into two fragments: one containing the QI, and the other containing the sensitive attribute. Each record in each fragment is then associated with the identifier of the group to which it belongs, to permit to loosely associate (satisfying ℓ-diversity) groups of quasi-identifiers

SSN	Name	DoB	Sex	ZIP	ID
		1960/03/12	M	98302	1
		1960/07/30	M	98389	1
		1960/07/12	M	98312	1
		1950/09/10	F	97401	2
		1950/09/04	F	97467	2
		1950/09/11	F	97434	2
		1960/03/20	M	98245	3
		1960/03/12	M	98290	3
		1960/07/25	M	98223	3

ID	Disease	Count
1	Flu	1
1	Cough	1
1	Gastritis	1
2	Stroke	1
2	Colitis	1
2	Asthma	1
3	Flu	2
3	Cancer	1

Figure 3: An example of a 3-diverse version of the dataset in Figure 1(a) with the Anatomy approach.

to groups of sensitive values. The fragment of the sensitive attribute also includes an additional attribute reporting the number of occurrences of each distinct value in each equivalence class, which is then only reported once. The datasets in Figure 3 illustrates an example of a 2-diverse version of the medical dataset in Figure 1(a) (again, after the removal of the last record), computed through the Anatomy approach. It is easy to see that the protection guarantees offered are the same as those of the 2-diverse dataset in Figure 2(b) computed through generalization: each respondent in the left-hand-side fragment can in fact be associated with at least ℓ=2 different sensitive values.

Data fragmentation can also be effectively adopted whenever the sensitive information to be protected is represented by generic associations among data items (attributes in the relational context), which may go beyond the traditional respondents' identity/sensitive information pair we have discussed so far. In this context, the original relation can be split in a set of fragments in such a way that no sensitive information is visible within the same fragment (e.g., [1, 8, 9]). The protection of such sensitive associations is then enforced by either restricting the visibility over some fragments, or by ensuring that fragments be unlinkable (i.e., no direct or indirect correlation can allow a recipient to reconstruct a single record that has been split by fragmentation) [12]. Fragments are then enriched with *loose associations* [13].

Similarly to Anatomy, records in the fragments are partitioned in groups of a desired cardinality, and information about the associations among groups are then released along with the fragmentation. Indeed, to guarantee an adequate protection degree, the partitioning of the records should be performed carefully, to ensure heterogeneity of the records that might be reconstructed.

2.3 Semantic Approaches

Differently from syntactic protection approaches, semantic approaches are typically based on the controlled distortion of the original data, to blur the actual values. The protection guarantees are strong and mathematically founded, but the price to be paid comes in terms of sacrificing data truthfulness.

Differential privacy [16] and its extensions are probably the most famous approaches that pursue a semantic privacy definition. The aim of differential privacy is to ensure that the release of a dataset does not disclose sensitive information of *any* individual, be her data included in the dataset or not. To this end, differential privacy is designed to permit to discover statistics (or, more generally, properties) of a dataset in its entirety, while ensuring that the probability of discovering the values of a single individual does not substantially differ due to her inclusion in/exclusion from the dataset itself. Roughly speaking, the release of a differentially private dataset should not increase the probability that a recipient can correctly guess the actual values of a target respondent. More precisely, given two datasets T and T' differing only for one record, a randomized function $\mathcal{K}$ (typically, the release function) satisfies *ϵ-differential privacy* if and only if $P(\mathcal{K}(T) \in S) \leq \exp(\epsilon) \cdot P(\mathcal{K}(T') \in S)$, with S a subset of the outputs of $\mathcal{K}$ and ϵ a privacy parameter (clearly, the lower the value of ϵ, the greater the protection offered). This guarantees that, given a result in S for the evaluation of $\mathcal{K}$ over T, the probability of observing the same result over T' remains negligible. This means that the removal/insertion of one record from/to the dataset does not significantly affect the result of the evaluation of function $\mathcal{K}$, and thus that the impact that an individual has on the outcome of $\mathcal{K}$ remains negligible. A direct consequence of this privacy guarantee concerns the fact that the privacy of individuals cannot be compromised by the possible external knowledge that a recipient can have, since the release of a differentially private dataset guarantees a limited information gain for recipients.

Differential privacy can operate in two different scenarios. In the *interactive scenario*, the one for which it had first been proposed, differential privacy is used to protect the answers to queries posed against a dataset that could contain sensitive information. In the *non-interactive scenario*, a sanitized version of the dataset to be shared is computed. In the interactive scenario, queries are typically evaluated on the original (unprotected) dataset, and the query results are distorted for instance by adding, for numerical values, *random noise* [17]. The typical distribution considered for the random noise is the *Laplace distribution* $Lap(\Delta(f)/\epsilon)$ with probability density function $P(x) = exp(-|x|/b)/2b$, where $b = \Delta(f)/\epsilon$ and $\Delta(f)$ is the maximum difference between the query result evaluated over T and over T'. In the non-interactive scenario, a differentially private dataset is instead directly produced and released, typically based on the evaluation of (differentially private) histogram queries, that is, on counting the number of records having a given value in the data domain.

It is interesting to highlight a recent extension of differential privacy, which can be used to collect data from a set of individuals who do not fully trust for confidentiality the agent in charge of the data collection. This approach, called *local differential privacy*, is based on the randomization of the individual pieces of data directly at the respondents' side, that is, before being collected (e.g., [18]). More precisely, given two pieces of data x and x' and a randomized function $\mathcal{K}$, local differential privacy ensures that $P(\mathcal{K}(x) \in S) \leq \exp(\epsilon) \cdot P(\mathcal{K}(x') \in S)$, with S a subset of the outputs of $\mathcal{K}$ and ϵ the privacy parameter.

Local differential privacy then limits the knowledge gain that the that even the agent in charge of collecting the individual pieces of data of the respondents can obtain from the collection. It is then interesting to note that the error introduced in the randomization of the responses can then be deducted by the collector upon receiving all the answers from all the participants, to obtain a fair representation of the true counts without knowing the original responses from the participants.

3 INFORMATION CREDIBILITY

In the previous section, we addressed the problem of protecting the confidentiality of data to be shared among different parties in scenarios where the implicit assumption is that data are genuine and correctly generated/collected, and hence the focus is on *data privacy*. When, however, as often happens above all through the use of social media, the source is unknown or potentially unreliable, and the contents are mainly exchanged in an unstructured way, it is necessary to focus on the *potential* information diffused, to verify its degree of *credibility*.

3.1 Credibility in the Social Web

In the 'offline' world, users could rely on traditional forms of information verification, such as the presence of traditional media *intermediaries* such as experts, by considering their reputation, or trust them based on first-hand experiences. Nowadays, in the Social Web scenario, almost everyone can spread contents on social media in the form of *User-Generated Content* (UGG), almost without any traditional form of trusted control. This 'disintermediation' process [19] has lead, on the one hand, to the democratization of the information diffusion, but, on the other hand, to the spread of possible fake news and misinformation, which we all know well. We live, in fact, in a so-called 'post-truth' era, in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. This is partially due to the fact that, in social media, user-created networks can become real *echo chambers* [26], in which one point of view dominates all the others, the verification of the statements has usually no effect, and this allows the repetition of unverified statements without refutation. The echo chamber phenomenon is emphasized by the filtering algorithms that are the basis of social media in proposing information of interest: by suggesting personalized (information) items that consider different elements of the user profile, such as location, past click-behavior and search history, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles, the so-called *filter bubbles* [36].

In this context, it becomes essential to try to find automatic solutions that assist the users to get out of their filtering bubbles and become aware of the level of *credibility* of the information they come into contact with.

3.1.1 Offline Information Credibility. In the research field of communication, the notion of *credibility* has been investigated since ancient times. In fact, among the first works that have come down to us that discuss this concept, there are the *Phaedrus* by Plato, and the Aristotle's *Rhetoric*, both dating back to the 4th Century BC. Over the years, depending on the context, credibility has been in turn associated with believability, trustworthiness, perceived reliability, expertise, accuracy, and with numerous other concepts or combinations of them [41]. Research in information credibility assessment has gradually moved from traditional communication environments, characterized by interpersonal and persuasive communication, to mass communication and interactive mediated communication, with particular reference to online communication [33]. The research undertaken by Hovland *et al.* [25] constitutes the first systematic work about credibility and mass media, focusing in particular on information *source credibility*. Later on, Fogg and Tseng in [20] stated that credibility is a *perceived* quality of the *information receiver*, and it is composed of multiple dimensions. In this sense, the process of assessing credibility involves different *characteristics*, which can be connected to [41]:

(*i*) the *source* of information;
(*ii*) the *information* itself, i.e., its structure and its content;
(*iii*) the *media* used to diffuse information.

3.1.2 Online Information Credibility. Online, and in the Social Web in particular, information credibility assessment deals with the analysis of both UGC and their authors' characteristics [35], and the intrinsic nature of social media platforms [38]. Specifically, this means to take into account *credibility features* connected to:

(*i*) *users* in social media (i.e., the information sources);
(*ii*) their *User-Generated Content* (i.e., the information they diffuse);
(*iii*) the *social relationships* connecting the involved entities in a virtual community (i.e., the main characteristic of an online social networking system).

Even if credibility is a characteristic perceived by individuals, credibility assessments should not be up to users, especially in the online environment [33]. In fact, humans have limited cognitive capacities to effectively evaluate the information they receive, especially in situations where the complexity of the features to be taken into account increases [29]. Furthermore, in interacting communities, the users' credibility perceptions are easily influenced by *crowd consensus* [22, 23], possibly leading to some issues such as echo chambers as previously discussed.

For these reasons, there is nowadays the need of developing interfaces, tools or systems that are designed to help users in automatically or semi-automatically assess information credibility. In the next section, we illustrate the approaches that have been proposed so far to assess credible information in the Social Web.

3.2 Approaches to Credibility Assessment

Recent approaches to information credibility assessment mostly rely on *data-driven* approaches and *model-driven* approaches. In the first case, starting from available data, a bottom-up model is learned to identify credible information with respect to non-credible one. In the second case, some domain knowledge is available, which is used to build a top-down model to tackle the considered issue. Another classification that can be made of the approaches that deal with the evaluation of the credibility of online information concerns the fact of studying the *propagation* of (false) online information or the attempt to produce a *classification* or a *ranking* of information based on its credibility level.

The approaches that fall into the above-mentioned categories are used to solve various tasks related to the evaluation of the credibility of online information, such as, for example, *opinion spam detection*, *fake news detection*, and *credibility assessment of online medical information* [47]. Although each of these tasks has its own peculiarities, the general notions that remain valid for each of the aforementioned tasks will be explained below.

3.2.1 Information Propagation. In general, *propagation-based approaches* are mainly concerned with studying the influence that *social bots* have on the dissemination of (false) information or how

low-credibility information spreads over the social network structure [42]. These approaches usually rely on the graph representation of the social network, and often employ unsupervised learning algorithms to detect cliques of malicious users or the so called *burst*, i.e., sudden increase of the use of a particular set of keywords in a very short period of time. Propagation-based approaches can rely on some pre-computed credibility values (usually learned from a classifier) and study their spread over the social network structure. The study of these kinds of problems has a slightly different aim (even if it is closely related) with respect to the assessment of information credibility, which is more related to classification-based approaches.

3.2.2 Information Classification (and Ranking). In *classification-based approaches*, fall both data-driven methods that are based on the use of (mostly supervised) learning algorithms to classify in a binary way information items based on their credibility level, and model-driven approaches which are based on some prior *domain knowledge*. The model-driven scenario includes both the use of the Multi-Criteria Decision Making (MCDM) paradigm, and the use of Knowledge Bases and Semantic Web technologies.

Approaches based on credibility features. When considering machine learning techniques and the MCDM paradigm to assess information credibility, different characteristics, i.e., *credibility features*, connected to different entities related to the information to be evaluated in terms of credibility are taken into account. As illustrated in Section 3.1, these features are generally related to the users in the virtual community, the information items that are generated and diffused, and the virtual relationships among users and other entities in the community. For this reason, it is possible to provide the following *classification of features*:

- *Behavioral features*: they are related to the users generating and diffusing information. They can be extracted both from public Web data, e.g., user ID, time of posting, frequency of posting, etc., and private/internal Web data, e.g., IP and MAC addresses, time taking to post an information item, physical location of the user, presence of an image profile, etc.
- *Content-based features*: they are related to the textual content of the information item. They can be both lexical features such as word n-grams, part-of-speech, and other lexical attributes, and stylistic features, e.g., capturing content similarity, semantic inconsistency, etc.
- *Social (Graph-based)* features: they capture complex relationships among users, the information they diffuse, and other possible entities (e.g., products and stores) in the social network.

Data-driven Approaches. In the case of using data-driven approaches that employ well-known (supervised) machine learning techniques (e.g., SVM, Random Forests, etc.), a binary classification of information items with respect to their credibility is obtained by training a model over the considered set of features and one or more suitable dataset(s). These approaches are hence based on a feature extraction and selection phase, and are dependent on the availability of (unbiased) labelled data, which is not always the case, as illustrated in the literature [47]. Furthermore, some of the machine learning techniques proving to be effective in the considered research field, are often inscrutable by observers (thy have a 'black-box' behavior), making it difficult to evaluate the importance of distinct and/or interacting features. A possible solution would be the study of approaches based on the so-called *eXplainable Artificial Intelligence* (XAI), referring to methods and techniques in the application of AI technology such that the results of the solution can be understood by human experts [24].

Model-driven Approaches. Approaches for assessing credibility that allow the human being to understand the result obtained are those modeling the considered problem as a Multi-Criteria Decision Making problem, characterized by the presence of a set of *alternatives*, i.e., the information items to be evaluated in terms of credibility, and a set of *criteria*, i.e., the considered set of credibility features. In an MCDM problem, each alternative 'satisfies' each criterion to a certain extent, producing this way a performance score, i.e., the credibility score, one for each criterion associated with the alternative. For each alternative, the aggregation of these multiple credibility scores produces an overall performance score, i.e., an overall credibility score.

In such kind of modeling of the problem, we are usually aware of the features to be considered, and of the *importance* that each feature has in terms of credibility. Furthermore, we model *satisfaction functions* to transform the values of the features into credibility scores. Finally, we can select suitable *aggregation operators* to obtain the overall credibility scores. This way, having an overall credibility score associated with each information item, we can provide both:

- A *classification* of information into genuine/fake (selecting a suitable threshold) [40].
- A *ranking* of the information items based on their overall credibility score [46].

Different an numerous are the families of aggregation operators to be considered for tackling the problem, depending on the preferences of the decision maker [45]. Recently, an MCDM approach has been proposed that allows to model interacting features, by employing the Choquet integral [37]; in this work, and in general in MCDM approaches based on aggregation operators, it can be complex to define the model when the number of features increases.

Approaches based on Knowledge Bases. Another way to use prior domain knowledge to assess the credibility of online information is to refer to the use of Knowledge Bases [43]. In this context, we do not start from a knowledge of what are the characteristics of credibility associated with information, and their importance, but the information that is known to be credible is expressed in terms of *facts*. This type of approach is used in particular for *automated fact checking*, in situations where manual information credibility assessment is not feasible, e.g., the number of experts is too limited with respect to the amount of information to be verified; crowdsourcing-based information credibility assessment does not guarantee the credibility of assessors.

Technically speaking, a *fact* can be modeled as a (Subject, Predicate, Object) (SPO) triple extracted from the given information that well represents it. For example, the information "Giacomo Puccini is a composer" can be expressed via the (GiacomoPuccini, Profession, Composer) triple.

Facts must be processed and cleaned up (redundant, outdated, conflicting, unreliable or incomplete data are removed) to build a *Knowledge Base*, i.e., a set of SPO triples. A graph structure, known as the *knowledge graph*, can be used to represent the SPO triples in a Knowledge Base, where the entities (i.e., subjects or objects in SPO triples) are represented as nodes and relationships (i.e., predicates in SPO triples) are represented as edges. Knowledge Bases are suitable candidates for providing *ground truth* to information credibility assessment studies, i.e., we can reasonably assume the existing triples in a Knowledge Base represent true facts.

Making these premises, to evaluate the credibility of to-be-verified information items, in turn represented as SPO triples, it is sufficient to compare them with the SPO triples contained in the Knowledge Base(s). Open issues of this kind of approach concern how to consider missing facts in the Knowledge Base(s), and, connected to this problem, how to constantly update the Knowledge Base(s) with up-to-date credible information.

4 DISCUSSION AND CONCLUSIONS

The advancements of ICTs make every day easier to collect, share, spread, and analyze huge amounts of data and information, often in a collaborative manner through Web 2.0 technologies. However, such a complex online scenario brings with it several concerns that should be properly addressed to ensure a profitable experience of online ecosystems that can be developed on top of it.

In this paper, we have discussed the main issues related to guaranteeing adequate protection to personal and/or sensitive data in the data sharing context, where data are supposed to be genuine and users trustworthy. Furthermore, we have illustrated the information credibility assessment research issue in the social media context, which is characterized by unknown or unreliable information sources and by the absence of traditional trusted intermediaries.

4.1 Open Issues and Further Research

While there have been major efforts by the research community to develop effective approaches for both the issues we have addressed, there is still a long way to go to advance research and develop novel solutions, as it is briefly summarized and discussed in the subsections below.

Data Protection. With reference to data protection, there has been a major debate on which, among *syntactic* and *semantic privacy definitions* and *approaches*, have to be considered as the 'right' ones.

Both approaches and definitions have their pros and cons: on the one hand, *syntactic approaches* enjoy the benefits of clear semantics and comprehensible protection guarantees, but address parts of the problem and build on some assumptions (e.g., the precise definition of the quasi-identifier) that can open the door to vulnerabilities; on the other hand, *semantic approaches* can offer stronger protection guarantees, but their semantics is unclear, and the distortion applied to datasets can reduce the utility for final recipients. Recent studies pointed out that there is actually room for both of them, possibly jointly adopted [10, 44].

Open research challenges in this context, then, can include the development of approaches combining the strengths of both approaches, the support for fine-grained and personal privacy preferences, the definition of privacy-preserving analytics, and the support for users in choosing the right privacy-preserving parameters (e.g., good values for k, ℓ, and ϵ in the approaches we have discussed).

Information Credibility. Numerous approaches belonging to distinct categories have been proposed up to now to assess information credibility. Every category presents both advantages and drawbacks: *propagation-based approaches* allows to effectively identify spam bots, but are affected by the problem of analyzing complex structures such as graphs, and, in some cases, by the need of having pre-computed credibility values to see how (genuine or fake) information spreads over the network; *classification-based approaches* are affected by inscrutability of results and data dependency in case of supervised black-box approaches, an by the difficulty of managing the complexity of the model when the number of criteria increases in an MCDM scenario; despite this, both turned out to be particularly effective in well defined tasks [47]. In approaches employing Knowledge Bases, some issues emerge about the treatment of missing information, conflict resolution and triples update in the knowledge graph(s), even if they can be particularly useful in automated fact checking.

Further research in the information credibility assessment field must deal with the above-mentioned issues, by developing, for example, hybrid approaches able to simultaneously exploit domain knowledge, model-driven aspects, and supervised learning when unbiased training data are available, even of a small number (e.g., labeled via crowdsourcing platforms by some experts).

4.2 Confidentiality and Credibility

To the benefits of online ecosystems and of our society at large, the approaches we have discussed in this paper could be mutually beneficial to each other, even if, at a first glance, the two research issues might seem orthogonal.

On the one hand, a challenge would be the consideration of confidentiality in the assessment of information credibility in the Social Web. As a matter of fact, personal/sensitive data of users can prove useful for the verification of information credibility (in particular, as aspects of the credibility of the source). Still, as illustrated in this paper, such data should be adequately protected. How can we avoid exposing personal/sensitive data in a scenario that already prompts users to do so, via *self-disclosure* [27], while being able at the same time to use these data to assess the credibility of the disseminated information? Actually, having effective means for protecting personal data could undoubtedly be useful for developing novel and improved information credibility assessment tools. Such tools could in fact operate on privacy-preserving yet truthful and useful sanitized versions of the personal data from the UGC collected from social media, to obtain stronger assessments while complying with data protection regulations. Furthermore, on these bases, trading anonymity for credibility in a controlled way could be a strategy to improve the assessment of the credibility of the information disseminated. This could be addressed by building, for example, a *trust and reputation system* through the monitoring and management of inter-actions across entities on social networks [6, 7]. In this system, rewards in terms of a higher credibility depending on the amount of (sanitized) personal data released on the trusted network could be provided [11].

On the other hand, also the concept of credibility could be useful in the data sharing context with respect to the confidentiality aspect. Actually, we note that credibility has been described in the literature as a characteristic of data *quality* [2]. Clearly, the higher the quality of a dataset, the better the utility for a recipient. Considering that data protection approaches (especially semantic ones) perturb the data, and that the amount of perturbation might be quite large, having effective means for assessing information credibility could be useful for evaluating the quality of a sanitized dataset, and, ultimately, for the definition of novel utility-aware data protection approaches.

At the end of this discussion, we can affirm that devising uniform solutions blending data protection and information credibility seems therefore a promising, although in a purely embryonic state, future research direction.

5 ACKNOWLEDGMENTS

This work was partly supported by the EC within the H2020 Program under grant agreement 825333 (MOSAICrOWN).

This paper is based on joint works by Giovanni Livraga and Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati, and on joint works by Marco Viviani and Gabriella Pasi, whom the authors would like to thank.

REFERENCES

[1] G. Aggarwal, M. Bawa, P. Ganesan, H. Garcia-Molina, K. Kenthapadi, R. Motwani, U. Srivastava, D. Thomas, and Y. Xu. Two can keep a secret: A distributed architecture for secure database services. In *Proc. of CIDR '05*, pages 186–199, Asilomar, CA, USA, Jan 2005. CIDR.

[2] C. Batini, M. Scannapieco, et al. Data and information quality. *Cham, Switzerland: Springer International Publishing*, 2016.

[3] C. J. Bennett. The european general data protection regulation: An instrument for the globalization of privacy standards? *Inf. Polity*, 23(2):239–246, 2018.

[4] G. Briscoe. Complex adaptive digital ecosystems. In *Proc. of MEDES'10*, pages 39–46, Bangkok, Thailand, Oct 2010. ACM.

[5] G. Briscoe and P. De Wilde. Digital ecosystems: evolving service-orientated architectures. In *Proc. of BIONETICS'06*, page 17, Trento, Italy, Dec 2006. ACM.

[6] B. Carminati, E. Ferrari, and M. Viviani. A multi-dimensional and event-based model for trust computation in the Social Web. In *Proc. of SocInfo'12*, pages 323–336, Lausanne, Switzerland, Dec 2012. Springer.

[7] P. Ceravolo, E. Damiani, and M. Viviani. Adding a peer-to-peer trust layer to metadata generators. In *Proc. of OTM'05 Workshops*, pages 809–815, Agia Napa, Cyprus, Oct 2005. Springer.

[8] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati. Combining fragmentation and encryption to protect privacy in data storage. *ACM Transactions on Information and System Security (TISSEC)*, 13(3):22:1–22:33, 2010.

[9] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Enforcing confidentiality and data visibility constraints: An OBDD approach. In *Proc. of DBSec '11*, pages 44–59, Richmond, VA, USA, Jul 2011. Springer.

[10] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. *Transactions on Data Privacy*, 6(2):161–183, 2013.

[11] E. Damiani and M. Viviani. Trading anonymity for influence in open communities voting schemata. In *Proc. of SocInfo'09*, pages 63–67, Warsaw, Poland, Jun 2009. IEEE.

[12] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. Fragmentation in presence of data dependencies. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 11(6):510–523, 2014.

[13] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. Loose associations to increase utility in data publishing. *Journal of Computer Security (JCS)*, 23(1):59–88, 2015.

[14] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Data privacy: Definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKBS)*, 20(6):793–817, 2012.

[15] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Empowering owners with control in digital data markets. In *Proc. of CLOUD'19*, pages 321–328, Milan, Italy, Jul 2019. IEEE.

[16] C. Dwork. Differential privacy. In *Proc. of ICALP '06*, pages 1–12, Venice, Italy, Jul 2006. Springer.

[17] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC '06*, pages 265–284, New York, NY, USA, Mar 2006. Springer.

[18] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proc. of CCS'14*, pages 1054–1067, Scottsdale, AZ, USA, Nov 2014. ACM.

[19] G. Eysenbach. Credibility of health information and digital media: New perspectives and implications for youth. In *Digital Media, Youth, and Credibility*, pages 123–154. The MIT Press, 2008.

[20] B. J. Fogg and H. Tseng. The elements of computer credibility. In *Proc. of CHI'99*, pages 80–87, Pittsburgh, PA, USA, May 1999. ACM.

[21] P. Golle. Revisiting the uniqueness of simple demographics in the us population. In *Proc. of WPES '06*, pages 77–80, Alexandria, VA, USA, Oct 2006. ACM.

[22] M. N. Hajli. Developing online health communities through digital media. *International Journal of Information Management*, 34(2):311–314, 2014.

[23] M. N. Hajli, J. Sims, M. Featherman, and P. E. Love. Credibility of information in online communities. *Journal of Strategic Marketing*, 23(3):238–253, 2015.

[24] A. Holzinger. From machine learning to explainable AI. In *Proc. of DISA'18*, pages 55–66, Kosice, Slovakia, Aug 2018. IEEE.

[25] C. I. Hovland, I. L. Janis, and H. H. Kelley. *Communication and persuasion.* New Haven: Yale University Press, 1953.

[26] K. H. Jamieson and J. N. Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press, 2008.

[27] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[28] F. Karakas. Welcome to world 2.0: the new digital ecosystem. *Journal of Business Strategy*, 30(4):23–30, 2009.

[29] A. Lang. The limited capacity model of mediated message processing. *Journal of Communication*, 50(1):46–70, 2000.

[30] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proc. of ICDE '07*, pages 106–115, Istanbul, Turkey, Apr 2007. IEEE.

[31] G. Livraga. Privacy in microdata release: Challenges, techniques, and approaches. In N. Crato and P. Paruolo, editors, *Data-Driven Policy Impact Evaluation: How Microdata is Transforming Policy Design*. Springer, 2018.

[32] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM TKDD*, 1(1):3:1–3:52, 2007.

[33] M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the Association for Information Science and Technology*, 58(13):2078–2091, 2007.

[34] M. J. Metzger and A. J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *J. of Pragmatics*, 59:210–220, 2013.

[35] M.-F. Moens, J. Li, and T.-S. Chua, editors. *Mining User Generated Content.* Social Media and Social Computing. Chapman and Hall/CRC, 2014.

[36] E. Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011.

[37] G. Pasi, M. Viviani, and A. Carton. A Multi-Criteria Decision Making approach based on the Choquet integral for assessing the credibility of User-Generated Content. *Information Sciences*, 503:574–588, 2019.

[38] L. Safko. *The Social Media Bible: Tactics, Tools, and Strategies for Business Success.* Wiley Publishing, 2nd edition, 2010.

[39] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6):1010–1027, 2001.

[40] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[41] C. C. Self. Credibility. In *An integrated approach to communication theory and research*, pages 449–470. Routledge, 2014.

[42] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787, 2018.

[43] B. Shi and T. Weninger. Fact checking in heterogeneous information networks. In *Proc. of WWW'16*, pages 101–102, Montréal, Québec, Canada, Apr 2016. International World Wide Web Conferences Steering Committee.

[44] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5):771–794, 2014.

[45] M. Viviani and G. Pasi. A multi-criteria decision making approach for the assessment of information credibility in social media. In *Proc. of WILF'16*, pages 197–207, Naples, Italy, Dec 2016. Springer.

[46] M. Viviani and G. Pasi. Quantifier guided aggregation for the veracity assessment of online reviews. *International Journal of Intelligent Systems*, 32(5):481–501, 2016.

[47] M. Viviani and G. Pasi. Credibility in Social Media: Opinions, News, and Health Information - A Survey. *WIREs Data Mining and Knowledge Discovery*, 7(5), 2017.

[48] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proc. of VLDB'06*, pages 139–150, Seoul, Korea, Sep 2006. VLDB Endowment.