# Co-utile Disclosure of Private Data in Social Networks

David Sánchez, Josep Domingo-Ferrer and Sergio Martínez

*UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics*
*Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia*
{*david.sanchez,josep.domingo,sergio.martinezl*}*@urv.cat*

## Abstract

Social networks (SNs) have become the mainstream web service by which users publish and share information. However, since much of that information is personal and sensitive, disclosing it in an uncontrolled way entails serious privacy risks. Paradoxically, most SNs assume that their users are willing to disclose sensitive data to others (even strangers) with little to no control. In this paper, we formalize the utility that rational users derive from participating in SNs, and argue that the current information exchange model is hardly sustainable from a rational viewpoint; actually, it goes against the interests of privacy-aware users. To tackle this issue, we propose several *co-utile* protocols for exchanging (sensitive) information among SN users. An interaction is said to be co-utile if the best way for a participant to increase her own utility is to help other participants increase theirs; hence, co-utile information exchange is self-enforcing and mutually beneficial for rational users. In this way, we ensure the sustainability of SNs in the long term, especially SNs with a sensitive scope (e.g., healthcare).

*Keywords:* Social networks, Data privacy, Information exchange, Co-utility

## 1. Introduction

Social Networks (SNs) have ushered in a new information sharing paradigm whereby the information published on the Internet is no longer generic or anonymous, but associated to individuals. In this respect, the Consumer Reports' 2010 State of the Net analysis [13] highlights that more than half of SN users usually share private information and, as a result, they are exposed to a number of privacy-related threats, such as spamming, phishing [50], discrimination (*e.g.*, in job application) [38] or bullying [15]. Studies have also shown that the increasing awareness of the privacy threats underlying personal data publication in SNs has negatively affected the information posting rate of SN users and that many users have shifted from posting to reading [40, 24]. Privacy-aware users constitute a significant problem for the sustainability of SNs [25], because SN "free-riding" (*i.e.*, getting information about others without offering information about themselves) may result in a functional standstill where no new information is offered on the SN.

Even though the consequences of the privacy concerns of SN users may affect any network, it is especially relevant in SNs with particularly sensitive scopes. For example, in LinkedIn, users disclose their CVs and their detailed professional experience for professional networking and to attract job offers or gain professional contacts; in healthcare-oriented SNs, such as PatientsLikeMe, patients share their medical conditions and healthcare experiences because they expect to learn from others' experiences. In this kind of SNs, privacy concerns may outweigh the utility benefits of information sharing and, thus, compromise the sustainability of the network in the long run.

SN users face two types of privacy risks: (i) those derived from the fact that all their data are hosted by a centralized SN operator, who may exploit and/or resell their personal data for business purposes, such as identity verification, marketing or personal profiling [46]; and (ii) those caused by releasing sensitive data in an uncontrolled way to other users in the SN, who may use these data for malicious purposes (*e.g.*, phishing, blackmailing, discrimination, etc.).

The former risk can be tackled by using distributed SN architectures, which avoid relying on centralized operators. Diaspora is probably the best-known example of a decentralized SN [47], even though other systems have also been proposed in the literature [14, 30]. Decentralized SNs allow users to install and manage their own personal web server that locally stores all their data (*e.g.*, photos, videos, etc.). Since users control their own data, they retain full ownership over the shared content, which is not subject to changing privacy policies or sellouts to third parties.

To mitigate the second privacy risk, in this paper we propose information exchange protocols that assist users in making informed rational decisions on what (sensitive) data they reveal to their peers in the SN. To ensure the sustainability of the SN, our protocols are grounded in the notion of *co-utility* [21, 20]. Specifically, *co-utile* protocols are those in which helping other peers increase their utilities is also the best way to increase one's own utility. In those SNs where the main utility is the information the users gain from others, we envision co-utile information exchange as a *quid pro quo* interaction whereby users only disclose their sensitive data to other peers that also disclose a similar amount of their own sensitive data. In this way, we aim at making users aware of the privacy risks inherent to disclosing their data (because data are characterized and exchanged according to their sensitivity), and at balancing the reciprocal disclosure of sensitive information caused by the information exchange, thus avoiding SN free-riding. Our protocols make sensitive data release compatible with disclosure control, thereby contributing to mitigating the privacy concerns of privacy-aware users, who are especially important in SNs with sensitive scopes. Moreover, since users are rationally motivated to contribute their own private data to the network to an extent sufficient to match what they obtain from the other peers, data release becomes self-enforcing and mutually beneficial for the involved users, and sustainable in the long term.

To attain the goals above, in this work:

- We characterize the utility a user derives from participating in the SN as a function of the information she obtains about other users and the privacy risk she incurs by disclosing her own data to others. The quantification of this utility relies on an automatic assessment of the privacy risks associated to the data the SN users may disclose (*e.g.*, profile attributes, messages, etc.), which are classified according to the sensitive topic to which they refer (*e.g.*, healthcare, religion, etc.).

- We use the previous characterization to design decentralized and *co-utile* information exchange protocols, which ensure that rational users (even purely selfish ones) will follow them; that is, our protocols motivate rational users to contribute to the SN and, therefore, they thwart free-riding and ensure the sustainability of the network.

- We mitigate the reluctance of users to disclose sensitive information to others by incorporating an also decentralized and co-utile reputation system. In this way, users can build *trust* in each other, while reputation makes them accountable for their behavior. The use of reputations also makes the information exchange between peers more efficient and straightforward.

- We propose extensions to our protocols to: i) support many-to-one information exchange (*e.g.*, within SN groups), and ii) normalize the disclosure risk assessment (to adjust the flow of exchanged information) when applying the protocols to users with significantly different levels of social exposure (and, thus, of privacy requirements).

The rest of the paper is organized as follows. Section 2 discusses related works proposing privacy-preserving mechanisms for SNs. Section 3 presents an automatic method to measure the utility a user derives from participating in the SN, as a function of the functionality (information) she obtains from her peers and the privacy risk she incurs when disclosing sensitive data. Section 4 provides background on co-utility and proposes two co-utile information exchange protocols for SNs: a basic one-to-one iterative and incremental information exchange mechanism, and another mechanism that relies on the reputation of users. Section 5 reports the results of several experiments carried out on synthetic users and highly sensitive (health) data. Section 6 describes protocol extensions for many-to-one information exchange and for exchange between users with asymmetric social exposure. The final section gathers conclusions and identifies some lines of future research.

## 2. Related work

To control the disclosure of sensitive data of SN users, social network operators (such as Twitter or Facebook) have implemented basic privacy settings that enable users to specify who may access certain data, such as their profile attributes or messages. More sophisticated approaches employ privacy policies, such as contracts, which specify who can access a certain resource [8, 10, 17].

However, the use of manually defined privacy settings/policies has been criticized because: (i) they are burdensome to manage and, as a result, most users seldom change the default settings, that generally make most user information public [41]; and (ii) users find difficulties to assess the privacy risks caused by disclosure of their data, whereas such an assessment is needed to define access control rules [49].

Regarding this latter issue, many authors have proposed mechanisms to assess the privacy risks inherent to users' data in SNs. In [28], a privacy risk score was presented to quantify the privacy risks caused by disclosing profile attributes. Attributes (*e.g.*, country, political views, religion) are associated a sensitivity value (*i.e.*, how embarrassing it is for a user to reveal an attribute to a certain other user). The privacy score is then calculated as the sum of attributes (weighted by their respective sensitivity) that are visible to all the peers in the network. Similar approaches with *ad hoc* privacy scores have been also contributed in [39, 49]. In [5, 42] the authors propose mechanisms that infer hidden attribute profiles of a user from the publicly available attributes of her friends in the SN. The number of attributes that can be inferred is used to calculate a privacy score. In [7] the authors introduce a cryptographic protocol to unlink the identities of the users when submitting *Likes* to an SN (because *Likes* may reveal some of the users' attributes). The authors also rely on distributed SNs to avoid a centralized provider that learns the sensitive data of users. In [26], the authors analyze the privacy issues that derive from tags submitted by the users, which may reflect their preferences and personal features. In [45], the authors describe a measure to evaluate the privacy of social graphs, that is, graphs that depict the social innerconnections between SN users.

The papers reviewed thus far only focus on the privacy risks related to attribute profiles and social connections, and neglect the risks inherent to textual messages, which account for most of the content currently released in SNs [48, 37].

In [39] the authors measure the privacy risks of user publications by relying on a manual association of their contents with attribute profiles. Then, the privacy risk is measured according to the information distribution of the message contents within a subset of social network users. In [16] the authors use structured knowledge bases (ontologies) to define disclosure thresholds for sensitive topics (*e.g.*, healthcare, religion, etc.). According to these, the terms appearing in textual publications that are more specific than the thresholds are considered risky, and access control rules are defined on this basis.

All the former methods require manually defining the sensitivity of the published contents with different degrees of granularity and, in some cases, with respect to the type of users the data may be disclosed to. This constitutes a significant burden on SN users, who may also lack the technical knowledge and awareness of the risks inherent to the data they publish. Moreover, the fact that each user may understand (and state) privacy risks in a different (and subjective) way makes it difficult to fairly compare the level of disclosure incurred by different users. In [37, 48] a more general and automatic approach based on information theory is presented. This mechanism identifies individual terms within the published contents and uniformly measures the privacy risk they entail according to their informativeness. Even though this mechanism yields an objective measurement of privacy risks, it systematically considers highly informative terms risky, regardless of whether they actually refer to a sensitive topic or not.

## 3. A privacy-functionality score

The cornerstone of the rational information exchange protocols we propose is a model of the utility that a privacy-concerned user derives from participating in an SN. In this section, we describe how to assess this dimension as the functionality the user obtains (*i.e.*, the sensitive information she gathers about her peers) divided by the privacy risk she incurs by revealing her own sensitive data to others. The basic idea of modeling a user's SN utility as the ratio between the information learned and the information disclosed by the user was first proposed in [18], but here we introduce more accurate ways of evaluating how informative is what is learned and how risky for privacy is what is disclosed. Moreover, whereas in [18] the disclosed information was limited to profile attributes, in this work we extend it to the whole user's data. Based on the resulting ratio, the protocols we propose will enable a rational (balanced) exchange of sensitive information among peers.

### *3.1. Measuring privacy risks of SN data*

Unlike other approaches in the literature [28, 39, 49, 5, 42], our assessment of privacy risks will account for *all* the data a user may release in an SN (*i.e.*, profile attributes, text messages and tagged multimedia files), regardless of their type and structure. Because these data mainly consist of unstructured text or textually tagged multimedia files, they are usually understood and analyzed by stakeholders (*i.e.*, content publishers, readers, SN operators and also potential attackers) according to their underlying semantics. Thus, the privacy risks incurred by a user disclosing these data (and, also, the utility her peers derive from the disclosed data) should be *semantically* quantified [1, 37]. Moreover, unlike most related works discussed in Section 2, our approach will be generic and *automatic*, hence overcoming the burden and subjectivity inherent to manual assessments of privacy risks.

To do so, we first evaluate whether each data piece (*e.g.*, an attribute value, a tag, (part of) a message, etc.) refers to a sensitive topic that should be protected against disclosure (*e.g.*, health, sexuality, ethnicity, etc.). Then, we measure the privacy risk of releasing the data piece on a sensitive topic, according to the amount of semantics the piece conveys. This approach is coherent with current research on textual data protection [1, 35], which assumes that pieces of data conveying a large amount of semantics are potentially risky, because they are the ones from which third parties gain most knowledge on the individuals.

However, measuring data semantics is not trivial because they are inherently human and qualitative features. Following the state of the art on textual data analysis [32] and document sanitization [33, 35, 2], we adopt an information-theoretic quantification of data semantics: the semantics conveyed by a textual term $t$ is quantified as the informativeness of the term (*i.e.*, its *Information Content* (IC)) computed as the inverse of the probability of occurrence of $t$ in corpora [31], that is

$$IC(t) = -\log(p(t)), \tag{1}$$

where, for the sake of concreteness, we take the logarithm to be binary. According to Expression (1), general terms (*e.g.*, *disease*) are considered less informative (and thus, potentially less sensitive) than specific ones (*e.g.*, *breast cancer*), because the former are more likely to appear in a text.

Under the same framework, the amount of information (semantics) that a term $t$ discloses about a certain topic $\tau$ can be measured by their mutual overlap of information (*i.e.*, their *Point-wise Mutual Information* (PMI)), which is computed as the ratio between their joint and marginal distributions [12]:

$$PMI(t;\tau) = \log \frac{p(t \cap \tau)}{p(t)p(\tau)}.$$

By applying these notions to our setting, and assuming that $t$ is a piece of data describing a feature of a user and $\tau$ is a sensitive topic for which disclosure should be controlled (*e.g.*, health, race, etc.), we have the following upper and lower bounds for their PMI [34]:

- If $t$ and $\tau$ are independent, that is, they co-occur in corpora just by chance, then $PMI(t;\tau) = 0$. In this case, we have that $t$ is not disclosing anything about $\tau$ and, hence, the privacy risk $t$ may cause on the sensitive topic $\tau$ is zero. Strictly speaking, PMI may also take negative values if $t$ and $\tau$ are exclusive: when they never co-occur, $PMI = -\infty$. However, since textual entities are in general correlated up to some degree [27], we can attribute rare co-occurrence or no co-occurrence to data sparseness in the probability calculus (*i.e.*, to the fact that not enough data are available to extract reliable conclusions from their analysis) rather than to real exclusiveness [34]. Therefore, for the sake of semantic coherence, we truncate negative PMI values to 0; that is, we use $PMI'(t;\tau) = \max(0, PMI(t;\tau))$.

- If $t$ and $\tau$ are perfectly correlated (they always co-occur), either for an occurrence of $t$ and/or an occurrence of $\tau$, PMI is maximized to $PMI(t;\tau) = -\log(p(t)) = IC(t)$ if $p(t \cap \tau) = p(\tau)$, or to $PMI(t;\tau) = -\log(p(\tau)) = IC(\tau)$ if $p(t \cap \tau) = p(t)$. In particular, if $PMI(t;\tau) = IC(\tau)$, we have that $t$ is providing information that fully refers to the sensitive topic $\tau$ (*e.g.,* $t$ might be a specialization of $\tau$), and the specific amount of information (semantics) $t$ discloses on $\tau$ is $IC(t)$.

Therefore, in case $PMI'(t;\tau) = IC(\tau)$, we adopt $IC(t)$ as the privacy risk of disclosing $t$ to peers in the $SN$ regarding $\tau$, that is, $PR_\tau(t)$. In practice, closely (even though not perfectly) correlated terms may also be risky if most of the information they reveal conveys $\tau$. To also consider this situation and to make the risk assessment more flexible, we incorporate a parameter $\alpha \in [0..1]$, which defines the relative amount of $\tau$'s information conveyed by $t$ that is considered risky. Formally, the $PR$ expression is defined as follows:

$$PR_{\tau}(t) = \begin{cases} IC(t) & if PMI'(t;\tau) \geq \alpha \times IC(\tau), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We apply the former notions to measuring the privacy risk resulting from the data disclosed by the user in the SN as follows. Let $T^a = \{t_1^a, \ldots, t_n^a\}$ be the set of data pieces referring to the user $u_a$. As stated above, $t_i^a$ might be individual profile attributes, textual tags associated to multimedia files (*e.g.*, tagged photos) or (part of) messages to be posted in the SN wall. Which data are considered and how they are tokenized in pieces can be configured depending on the specific privacy needs or the particularities of the SN (more details are given in Section 5); for example, complete messages can be divided in sentences or noun phrases for a finer-grained assessment of privacy risks.

On the other hand, let $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ be the set of sensitive topics that may put the privacy of SN users at risk. One might think of individual users manually defining which topics they consider sensitive; however, to make the information exchange coherent among peers, it seems better to fix sensitive topics for all SN users according to one or both of the following criteria:

- The thematic scope of the SN (*e.g.*, *health* topics in medical-oriented SNs such as PatientLikeMe);

- The topics declared as private by applicable legal frameworks (*e.g.*, the EU Data Protection Act [23] defines data related to political opinions, race, sexual orientation, religion and health as private).

By applying Expression (2) to each pair $(t_i^a, \tau_k)$, where $t_i^a \in T^a$ and $\tau_k \in \mathcal{T}$, we identify the data pieces $t_i^a$ that entail a privacy risk for user $u_a$ and each sensitive topic $\tau_k$. Notice that each $t_i^a$ may entail risks for *several* $\tau_k$; for example, a *sexually transmitted disease* may disclose sensitive information about *health*, but also about *sexuality*.

Finally, we measure the *accumulated* privacy risk of users' data $T^a = \{t_1^a, \ldots, t_q^a\}$ for all $\tau_k \in \mathcal{T}$ as follows:

$$PR_{\mathcal{T}}(T^a) = \sum_{\forall \tau_k \in \mathcal{T}} \sum_{\forall t_i^a \in T^a} PR_{\tau_k}(t_i^a). \tag{3}$$

Notice that, as stated in Expression (2), the pieces in $T^a$ that do not convey the sensitive topic do not increase the privacy risk of the user; hence, we do not consider them risky regardless of their informativeness. This is a fundamental difference w.r.t. related works also relying on information-theoretic measures [48, 37], which systematically consider specific/informative terms risky, whether or not they actually refer to a sensitive topic.

**Proposition 1.** *For any disclosure of risky (sensitive) information by a user $u_a$, it holds that $PR_{\mathcal{T}}(T^a) > 1$.*

*Proof.* According to Expression (3), $PR_{\mathcal{T}}(T^a)$ is the sum of the privacy risks of data pieces in $T^a$. Now, by assumption, $T^a$ contains at least one data piece $t$ that conveys risky information for $u_a$. According to Expressions (1) and (2), the privacy risk of $t$ is $IC(t) = -\log(p(t))$. Since we assume binary logarithms, $IC(t) \leq 1$ only if $p(t) \geq 1/2$, that is, if the probability of occurrence of $t$ is at least 50% in corpora. Assuming that such a highly frequent $t$ exists, it cannot convey risky information, which is a contradiction. Hence, it has to be $IC(t) > 1$, and therefore $PR_{\mathcal{T}}(T^a) > 1$. □

### *3.2. A privacy-functionality score*

As introduced above, our definition of the utility the users derive from their participation in the SN can be summarized as the amount of sensitive information each user $u_a$ learns about her peers in the SN (*i.e.*, the functionality $u_a$ obtains) divided by the privacy risks $u_a$ incurs by revealing her own sensitive information to others. This "rational" utility may not probably explain the attitude of the typical Facebook user, who tends to disclose data just for social visibility [25], and without caring much about what she gets in return for her own privacy. However, our vision is more adapted to SNs with the sensitive scopes we mention above (*e.g.*, LinkedIn, PatientsLikeMe), where users rationally disclose their information (i) in a more targeted way (because, instead of just looking for social visibility, they expect to gain information in return), and (ii) with more caution (because they are aware of privacy risks due to the sensitivity of the data they disclose) [24].

In the sequel, we formalize this utility for pairwise interactions between two users $u_a$ and $u_b$:

- On the one hand, the privacy risk incurred by user $u_a$ when disclosing her data $T^a$ (or a subset) to a peer $u_b$ can be measured with the $PR$ score presented above (Expression (3)), which is the reciprocal of the privacy-preservation utility that concerns privacy-aware users: the greater $PR_{\mathcal{T}}(T^a)$, the lower is the privacy-preservation utility for user $u_a$.

- On the other hand, the functionality $u_a$ gets from $u_b$ can also be seen as the privacy risk ($PR$) incurred by $u_b$ when disclosing his own data to $u_a$. This perfectly fits our information-theoretic assessment of privacy risks, because the functionality $u_a$ obtains is the *amount of (sensitive) information* she learns from $u_b$.

Since disclosure risks are uniformly and objectively measured for all users (*i.e.*, $\mathcal{T}$ is fixed beforehand), we can coherently compare and integrate both the incurred privacy risks w.r.t. other peers and the functionality a user gets from those peers.

Formally, we quantify the utility a user $u_a$ derives from interacting with her peer $u_b$ in the SN by using the following privacy-functionality score

$$PRF(u_a, u_b) = \begin{cases} \frac{PR_{\mathcal{T}}(T^b)}{PR_{\mathcal{T}}(T^a)} & \text{if } PR_{\mathcal{T}}(T^a) > 1, \\ PR_{\mathcal{T}}(T^b) & \text{otherwise,} \end{cases} \quad (4)$$

where $T^x$ is the data user $x \in \{a, b\}$ discloses to the other user. By Proposition 1, the first case in Expression (4) corresponds to $u_a$ disclosing risky information, whereas the second case corresponds to $u_a$ disclosing either no information or at most non-risky information.

Our score has several advantages vs. other SN utility scores proposed in the literature [18]:

- It can be automatically assessed and updated as users release more data in the SN.
- Since it measures privacy and utility in an objective and uniform way, the incurred disclosure risks and gained information (functionality) can be fairly compared and integrated.
- The previous feature guarantees that self-interested attempts at tampering with or biasing privacy/functionality scores are ineffective, because peers can easily check the actual informativeness of the data they obtain from other peers. As we discuss in the following sections, this ensures co-utility in the rational information exchange protocols we propose.

According to Expression (4), $PRF$ decreases as the privacy risk of $u_a$'s disclosed data ($PR_{\mathcal{T}}(T^a)$ in the denominator) increases. Also, $PRF$ increases as $u_a$ has access to more (and more informative) sensitive data from $u_b$. Since $PRF$ is maximized when $u_a$ does not disclose any privacy-risky data (*i.e.*, $PR_{\mathcal{T}}(T^a) \leq 1$ and, thus, $PR_{\mathcal{T}}(T^b)$ is not divided by any factor), the dominant strategy for the users in a setting in which their utility is only defined by the privacy/functional dimensions in Expression (4) is to disclose no information or only trivial information to the SN; that is, to behave as free-riders. Therefore, if all users follow this strategy and the SN does not implement any mechanism to prevent it, the SN will shut down.

## 4. Co-utile protocols for rational content disclosure in SN

### *4.1. System model*

The approach we follow to tackle the rational free-riding behavior discussed above is based on i) decentralization of social network interactions by following a peer-to-peer model and ii) direct reciprocity (in other words, *quid pro quo*) of information disclosure. Specifically, the protocols we propose are strictly peer-to-peer and decentralized, and they are meant to be executed in distributed SN architectures. Such architectures support both one-to-one and many-to-one information exchanges, and users in them locally manage and store their own data (*e.g.*, photos, videos, etc.) in a personal web server. In this way, the users themselves are the ones who manage the information exchange without relying on a centralized SN operator. This forestalls privacy concerns that users might have vs. a central operator compiling their data (note that such concerns might deter users from releasing data to the SN).

Moreover, the protocols we propose make information release sustainable by relying on the notion of *co-utility*, which characterizes interactions in which the best alternative for rational users to increase their utilities is to help other users increase theirs. This mutually beneficial collaboration ensures that the protocols are self-enforcing and that peers do not find incentives to deviate from the protocols (*i.e.*, attack them or tamper with them).

### *4.2. Background on co-utility*

Co-utility models a kind of interaction between rational agents (*e.g.*, the users of an SN) in which the best option for each agent to reach her own goal (*e.g.*, to get information from the SN) is to help other agents reach theirs (*e.g.*, to help them get information from the SN, possibly by providing them with such information). Since we are dealing with rational agents, that is, agents that act strategically according to utility functions, game theory is a natural framework to formalize this concept. In [21, 20] we defined co-utility for scenarios that can be represented as perfect-information games; these are games in which each agent making a decision knows the utility payoffs of all agents under the various possible actions (or sequences of actions) they may execute, plus any previously made decisions. We represent these games in the so-called extensive form, which is a tree where: (i) nodes are the points where decisions are made, (ii) each node is labeled with the name of the agent (*e.g.*, SN user) making the decision, (iii) outgoing edges in a node represent the available choices (actions) at that node (*e.g.*, to exchange or not some data), and (iv) each leaf node is labeled with the tuple of utility payoffs that agents obtain when the node is reached.

By using this extensive form, we can view a *protocol* (*i.e.*, the actions needed for the completion of a task) as a path that traverses the tree representing the game.

Co-utility focuses on self-enforcing protocols, that are those from which agents have no rational incentive to deviate. That is, no agent can increase her utility by deviating from the protocol, provided that the other agents stick to it. In game-theoretic terms, this means that, at each successive node of the protocol path, sticking to the next action prescribed by the protocol (taking the next edge in the path) is an *equilibrium* of the remaining subgame (the subtree rooted at the current node), that is, a *subgame perfect equilibrium* of the game.

We say that a self-enforcing protocol is *co-utile* if it results in mutually beneficial collaboration (in terms of utility payoffs) between the participating agents. More specifically, a protocol $\mathcal{P}$ is co-utile if and only if *the three* following conditions hold:

1. $\mathcal{P}$ is self-enforcing;

2. The utility derived by each agent participating in $\mathcal{P}$ is strictly greater than the utility the agent would derive from not participating;
3. There is no alternative protocol $\mathcal{P}'$ giving greater utilities to all agents and a strictly greater utility to at least one agent.

The first condition ensures that, if participants engage in the protocol, they will not deviate. The second condition is needed to ensure that engaging in the protocol is attractive for everyone. The third condition can be rephrased in game-theoretic terms by saying that the protocol is a Pareto-optimal solution of the underlying game.

### *4.3. A co-utile protocol for sensitive information exchange*

The first protocol we propose implements a reciprocal, balanced and sequential exchange of (sensitive) information between SN users. Specifically, users no longer publish data in the SN without control, but decide which data will be (iteratively) disclosed to their peers. Being a strict P2P decentralized protocol, in principle it is designed for two users, $u_a$, $u_b$; nonetheless, in Section 6 we will detail how it can be extended to groups of users.

As discussed above, we assume that all users in the SN share the same set of sensitive topics, $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$, and that they implement the privacy risk assessment procedure we detailed in Section 3.1. Recall that $T^a = \{t^a_1, \ldots, t^a_n\}$ is the set of data pieces referring to user $u_a$. The protocol is defined as follows:

While simple, Protocol 1 has the shortcoming of requiring that the data pieces exchanged between $u_a$ and $u_b$ provide a similar amount of information on $\tau_k$. However, it may happen that $u_b$ only has data pieces $t^b_j \in T^b$ that are significantly less informative or more informative than the $t^a_*$ he received from $u_a$. In the former case, we can extend the protocol so that the exchange of data pieces is groupwise (*i.e.*, $u_b$ can release $S^b = \{t^b_1, \ldots, t^b_q\}$, where $S^b \subseteq T^b$ in response to $t^a_*$, if $PR_{\tau_k}(S^b) \approx PR_{\tau_k}(t^a_*)$). In the latter case, when $u_b$'s pieces are more informative than $u_a$'s, we can allow $u_b$ to release a generalization of $t^b_*$ (that is, $g(t^b_*)$) instead of $t^b_*$; by definition, a generalization of a term (*e.g.*, $g(AIDS) = \mathit{immunological\ disorder}$) discloses a *strict subset* of the semantics/information of the term (*i.e.*, $IC(\mathit{immunological\ disorder}) < IC(AIDS)$ and $PMI'(AIDS; \mathit{immunological\ disorder}) = IC(\mathit{immunological\ disorder})$).

Protocol 1 also assumes that the information that is iteratively exchanged between users does not overlap; that is, $PMI'(t^a_i; t^a_j) = 0, \forall t^a_i, t^a_j \in T^a$ and $PMI'(t^b_i; t^b_j) = 0, \forall t^b_i, t^b_j \in T^b$. However, since textual terms appearing in a context (*i.e.*, the SN account) are usually correlated [27], the data piece $t^a_*$ disclosed to $u_b$ at a certain iteration does not provide $IC(t^a_*)$ to $u_b$, but just $IC(t^a_*) - PMI'(t^a_*; \mathit{Received\_T}^b)$; for example, if $u_b$ already received from $u_a$ that the latter suffers from an *immunological disorder* due to *unprotected sexual intercourse*, then, if $u_a$ discloses to $u_b$ that she suffers from *AIDS*, the *new* information that $u_b$ is acquiring is not $IC(AIDS)$, but just $IC(AIDS) - PMI'(AIDS; \{\mathit{immunological\ disorder}, \mathit{unprotected\ sexual\ intercourse}\})$, because *AIDS* and $\{\mathit{immunological\ disorder,\ unprotected\ sexual\ intercourse}\}$ are closely

**Protocol 1** [Quid-pro-quo information exchange ($u_a$, $u_b$)]

$u_a$ and $u_b$ agree on a sensitive topic $\tau_k \in \mathcal{T}$ they are interested in about each other. Then $u_a$ does:

```
Set $Quit := 0$
Set $Disclosed\_T^a = \emptyset$
Set $Received\_T^b = \emptyset$
while $Quit = 0$ do:
    if $u_a$ has already disclosed all her data on $\tau_k$ to $u_b$
    ($i.e.$, $t_i^a \in Disclosed\_T^a$, $\forall t_i^a \in T^a$ such that $PR_{\tau_k}(t_i^a) > 1$), then
        set $Quit := 1$
    else
        Disclose to $u_b$ the data piece $t_*^a$ such that

            $$t_*^a = \arg \min_{t_i^a \in T^a \wedge t_i^a \notin Disclosed\_T^a \wedge PR_{\tau_k}(t_i^a) > 1} PR_{\tau_k}(t_i^a), \quad (5)$$

        that is, the least informative data piece that produces disclosure on $\tau_k$
        among those not yet disclosed to $u_b$.
        Add $t_*^a$ to $Disclosed\_T^a$.
    end if
    Request $u_b$ to disclose a $t_*^b$ such that $t_*^b \in T^b$, $t_*^b \notin Received\_T^b$ and
    $PR_{\tau_k}(t_*^b) \approx PR_{\tau_k}(t_*^a)$, that is, a data piece of $u_b$ that discloses a similar
    amount of information on $\tau_k$ as already disclosed by $t_*^a$.
    if $u_a$ does not receive $u_b$'s $t_*^b$ or $PR_{\tau_k}(t_*^b) \ll PR_{\tau_k}(t_*^a)$, then
        set $Quit := 1$
    else
        add $t_*^b$ to $Received\_T^b$.
    end if
end while
```

correlated and their mutual information is positive. Such informational overlaps can be accounted for by Protocol 1 by subtracting from the $PR$ of the disclosed/received data pieces that appear at lines 5, 11 and 12 (*i.e.*, $t_i^a$, $t_*^a$, $t_*^b$, respectively) their mutual information with the set of already disclosed/received data pieces (*i.e.*, $Disclosed\_T^a$, $Received\_T^b$, respectively); this means using $PR_{\tau_k}(t_i^a) - PMI'(t_i^a; Disclosed\_T^a)$, $PR_{\tau_k}(t_*^a) - PMI'(t_*^a; Disclosed\_T^a)$ and $PR_{\tau_k}(t_*^b) - PMI'(t_*^b; Received\_T^b)$ instead of $PR_{\tau_k}(t_i^a)$, $PR_{\tau_k}(t_*^a)$ and $PR_{\tau_k}(t_*^b)$, respectively.

Provided that the desire of learning information about peers (the numerator $PR_{\mathcal{T}}(T^b)$ in Expression (4)) is the dominant utility, Protocol 1 provides equal mutual benefits to the participants. Moreover, since the only way for a user to get data from peers is by disclosing a similar amount of her own data, the protocol is self-enforcing and prevents free-riding; thus Protocol 1 is co-utile. Also, given that an information exchange finishes when a user does not reciprocate the data received from her peer, users become motivated to contribute more and more of their own data if they want to learn a similar amount from others; this guarantees the sustainability of the SN.

However, Protocol 1 has the shortcoming that $u_a$, being the initiator, first takes the risk of not being reciprocated by $u_b$. What is more, by systematically refusing to reciprocate exchanges initiated by other peers, $u_b$ may learn small pieces of data from those peers "for free". Therefore, users may be reluctant to initiate the interaction because of the privacy risk it entails (the denominator $PR_{\mathcal{T}}(T^a)$ in Expression (4)); note, however, that the "loss" of $u_a$ is limited to the data piece disclosed in the unreciprocated iteration, which, by Expression (5), is the least informative one among those that could have been disclosed. In this case, co-utility may not hold and the information exchange within the SN may terminate. To tackle this issue, in the next section we mitigate the initiator's reluctance by leveraging user reputations and incorporating them into the information exchange protocol.

### *4.4. Reputation-based information exchange protocol*

The risk taken by the initiator $u_a$ can be mitigated by relying on past experiences with $u_b$, either direct interactions between $u_a$ and $u_b$ or interactions of $u_b$ with other peers in the SN. A natural way to capture and quantify the success of such past experiences is to use a *reputation system*. Reputation, which captures the opinion of the community on each peer, has at least two positive effects [19]:

- It allows users to build *trust*, which can neutralize the negative utilities related to mistrust. The higher a user's reputation, the more trusted she is by other peers.

- It makes users accountable for their behavior: if a user misbehaves (*e.g.*, he systematically refuses to provide his data after receiving those of others), his reputation worsens and his peers mistrust him more and more and become less and less interested to exchange information with him. In this manner, malicious agents (who may try to subvert the system, even irrationally) may be identified (via a low reputation) and penalized (*e.g.*, through limitation or denial of service).

Within our information exchange protocol, the reputation $s_a$ of a user $u_a$ in the SN can be understood as the disclosure she underwent in past interactions; that is, the reputation can be expressed in the same units as the actual information disclosure occurring during the information exchange. To manage and update reputations in a decentralized network, we need a reputation management protocol that is decentralized itself, such as the one we propose in [19]. Specifically, the protocol in [19] generalizes the well-known EigenTrust reputation calculation mechanism and introduces a more secure distributed calculation that cancels the benefits of deviating from (*e.g.*, tampering with) the protocol. Thanks to these modifications, the reputation calculation becomes scalable to large networks (because the calculation is distributed and parallelized among the users of the network), and robust against a number of classical attacks: self-promotion, whitewashing, slandering and denial of service [6]. In this way, the reputation management becomes *itself* co-utile and reputation can be seamlessly used as a mechanism to enforce co-utility in protocols in which negative utilities would otherwise rule it out [19], which is precisely the case of Protocol 1.

The basic idea of the reputation mechanism is to calculate a global reputation $s_a$ of a user $u_a$ based on aggregating the local opinions of the peers that have interacted with $u_a$. If we represent the local opinions by a matrix whose component $(i, j)$ contains the opinion of user $u_i$ on user $u_j$, the distributed calculation mechanism computes global reputation values that approximate the left principal eigenvector of this matrix. The interested reader can refer to [19] for a formal step-by-step description of the protocol, its scalable distributed calculation, its robustness to attacks and its co-utile nature.

Assuming that all users in the SN have global reputations $s_x$ computed from past experiences and that these reputations measure the disclosure users incurred in past interactions, we propose using them to mitigate the initiator's risk in Protocol 1, as follows.

Protocol 2 offers several properties as a result of incorporating $u_b$'s reputation:

- *Co-utility.* At line 05 ("if" clause), $u_a$ refuses to send any information to $u_b$ if $u_b$'s reputation (*i.e.*, the information disclosed by $u_b$ in past iterations) is lower than the disclosure level of $u_a$'s data pieces $t_i^a \in T^a$. That is, $u_a$ only *trusts* and discloses data to her peers if in previous iterations these peers have disclosed information at the same level $u_a$ has. We thereby mitigate the reluctance of users to initiate the information exchange, because trust cancels the fear by users of not being reciprocated, which is necessary to ensure the co-utility of the protocol.

- *Sustainability.* At line 08 ("else" clause), $u_a$ uses $u_b$'s reputation to decide how much information she is going to disclose, instead of just disclosing the least informative piece as in Protocol 1. In this way, users with high

**Protocol 2** [Reputation-based quid-pro-quo information exchange ($u_a$, $u_b$)]

$u_a$ and $u_b$ agree on a sensitive topic $\tau_k \in \mathcal{T}$ they are interested in. $u_a$ knows $u_b$'s global reputation in the SN, $s_b$.

```
Set $Quit := 0$
Set $Disclosed\_T^a = \emptyset$
Set $Received\_T^b = \emptyset$
**while** $Quit = 0$ **do**:
  **if** $u_a$ has already disclosed all her data on $\tau_k$ to $u_b$
  (*i.e.*, $t_i^a \in Disclosed\_T^a$, $\forall t_i^a \in T^a$ such that $PR_{\tau_k}(t_i^a) > 1$), or all her
  non-disclosed data are more informative on $\tau_k$ than $u_b$'s reputation $s_b$
  (*i.e.*, $\forall t_i^a \in T^a$ such that $t_i^a \notin Disclosed\_T^a$ it holds that $PR_{\tau_k}(t_i^a) > s_b$),
  **then**
    set $Quit := 1$
  **else**
    Disclose to $u_b$ the data piece $t_*^a$ such that $t_*^a \in T^a$, $t_*^a \notin Disclosed\_T^a$
     and $PR_{\tau_k}(t_*^a) \approx s_b$ that is, $u_a$ discloses a non-disclosed data piece that
     produces as much disclosure as $u_b$'s reputation.
    Add $t_*^a$ to $Disclosed\_T^a$.
  **end if**
  Request $u_b$ to disclose a $t_*^b$ such that $t_*^b \in T^b$, $t_*^b \notin Received\_T^b$ and
  $PR_{\tau_k}(t_*^b) \approx PR\tau_k(t_*^a)$, that is, a data piece of $u_b$ that discloses a similar
  amount of information on $\tau_k$ as already disclosed by $t_*^a$.
  **if** $u_a$ does not receive $u_b$'s $t_*^b$, **then**
    $\textsc{Update}(s_b, -PR_{\tau_k}(t_*^a))$
    set $Quit := 1$
  **else**
    $\textsc{Update}(s_b, (PR_{\tau_k}(t_*^b) - PR_{\tau_k}(t_*^a)))$
    add $t_*^b$ to $Received\_T^b$.
  **end if**
**end while**
```

reputations will receive further information faster than with Protocol 1. This will motivate users to maintain their reputations as high as possible, which can only be achieved by reciprocating information exchanges with *more* information. This reinforces the sustainability of the SN.

- *No free-riding.* If $u_b$ attempts free-riding on $u_a$ by not sending any information, $u_a$ punishes $u_b$ at line 13 ("if clause") by lowering $u_b$'s reputation as much as the unrequited information disclosure. This will cause $u_b$'s reputation to decrease and, thus, it will limit the chances of $u_b$ getting data from other peers. As a consequence, free-riding and protocol abuse are thwarted.

- *Fairness.* At line 16 ("else" clause), $u_a$ reflects the differential between the amount of disclosed/received information in $u_b$'s reputation; since the updated reputation will be considered in the next iteration to determine the amount of information to be exchanged, fairness is ensured.

Reputation can also be used to prevent malicious behaviors, such as users releasing sensitive but fake data in order to get information from others, or users releasing the confidential information they got from others to third parties. When these behaviors are detected, the affected users may punish malicious users by lowering their reputations. Therefore, if a user persists in malicious behaviors, his reputation will significantly decrease (or become zero), which will prevent him from getting new information. By raising awareness of this punishment, users will be rationally motivated to avoid malicious behaviors, which will ensure the co-utility of the information exchange.

In practice, once the set of sensitive topics and the criteria to evaluate their $PR$ (see Expression (2)) have been defined and agreed upon by the members of the SN, our protocols can operate in an automatic way, both during the assessment of disclosure risks and during the iterative exchange of sensitive data. However, if users wish to have more control on the information exchange, the protocols can be used just as *assistants* to the human user in deciding whether to disclose a piece of data to a peer; in this latter scenario, the final decision is made by the data owner, who may contradict the action proposed by the system (either because she does not agree with the automatic assessment of privacy risks or if utility components other than those modeled in Expression (4) influence her decision).

## 5. Empirical study

In this section, we report on experimental results obtained from simulating the two co-utile protocols proposed in the previous section. We created a pair of synthetic users that exchanged sensitive information about their medical conditions; thus, the sensitive topic was $\tau$ = *healthcare*. As usual in the field of document protection [11, 36, 33], we associated with each of the two users one disease considered as highly sensitive according to current legal frameworks [44] and we took the information on the disease from the Wikipedia; specifically, one user was assumed to suffer from *HIV* and the other from *Hepatitis*.

To implement the disclosure risk analysis detailed in Section 3.1 on the unstructured textual data of the users, we relied on natural language analysis. Sensitive entities are either concepts (*e.g.*, diseases) or instances (*e.g.*, hospital names), and those are represented in a text as nouns or noun phrases (NPs) [34]. Thus, we syntactically analyzed the text data to detect such noun phrases. This was done by means of a pipeline of natural language analyses consisting of sentence detection, tokenization of words, part-of-speech tagging to identify the grammatical categories of words, and syntactic parsing to identify phrases.

Once data pieces (NPs) had been detected, we assessed the privacy risk of each NP w.r.t. $\tau = healthcare$ by means of Expression (2) with $\alpha = 0.6$. Recall that $\alpha$ is a global parameter that sets the minimum degree of information overlap between each NP and $\tau$ to consider the former risky; this means that the NPs we considered risky in the experiments were those whose semantics mostly referred to healthcare topics (60% overlap or more). Notice that, since $\alpha$ was the same for all users, it only influenced the number of NPs considered risky and, thus, the number of protocol iterations, but not the tendencies and differences between the accumulated PR of the users, which were features we were interested in.

The information-theoretic calculations in Expression (2) require estimating the probability of (co-)occurrence of the NPs from corpora. To obtain representative assessments, it is crucial that the corpora be large and heterogeneous enough to capture the information distribution at a social scale. The Web, being the largest electronic repository available, fits those requirements and, in fact, has been broadly employed to estimate the distribution of the information in the fields of data semantics [4] and data privacy [11, 35, 36]. Specifically, to gather term probabilities of (co-)occurrence from the Web, we used the *hit count* provided by Google when querying the terms and term combinations of interest:

$$\begin{array}{l} p(t) = \frac{hits(''t'')}{total\_webs}, \\ p(t_i \cap t_j) = \frac{hits(''t_i'' AND ''t_j'')}{total\_webs}, \end{array} \tag{6}$$

where *total_webs* is the total number of web resources indexed by Google. Notice the use of quotation marks to constrain the search to exact matches of the query and the *AND* operator to look for co-occurrences of two (or more) terms.

The first experiment was carried out with Protocol 1 (Section 4.3) and the two synthetic users: one providing information about *HIV* (which we denote by $u^{HIV}$) and the other one about *Hepatitis* (denoted by $u^{HEP}$). As shown by the horizontal lines in Figure 1, the data corresponding to $u^{HIV}$ convey more information and, thus, entail more accumulated $PR$ (Expression (3)), than the data of $u^{HEP}$. Thus, releasing their information as done in a standard SN setting would result in significant disclosure risk asymmetry: $u^{HIV}$ would gain more sensitive information about $u^{HEP}$ than the latter about the former.

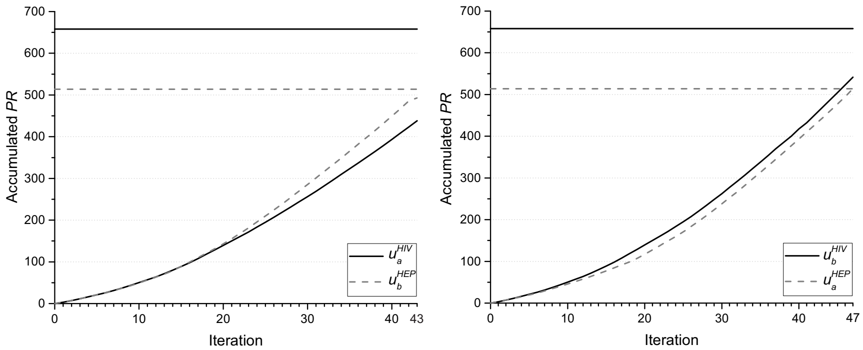

Figure 1: Protocol 1 between two users. Horizontal lines: accumulated $PR$ of the data sets held by the two users. Non-horizontal full lines: accumulated $PR$ for each user as a function of the number of iterations of Protocol 1. Left: the HIV user initiates the exchange, so that users are $u_a^{HIV}$ and $u_b^{HEP}$. Right: the HEP user initiates the exchange, so that users are $u_a^{HEP}$ and $u_b^{HIV}$.

Protocol 1 was executed in two different settings: (i) the user whose data had the largest accumulated $PR$ (user $u^{HIV}$) initiated the protocol (accordingly, in this setting we denoted users by $u_a^{HIV}$ and $u_b^{HEP}$), and (ii) the user with the lowest accumulated $PR$ (user $u^{HEP}$) initiated the protocol (accordingly, users were denoted by $u_a^{HEP}$ and $u_b^{HIV}$). For these tests, we only allowed the exchange of individual NPs, rather than groups of NPs. Figure 1 shows the results of the execution in both settings. The X-axis corresponds to the number of protocol iterations performed (iterations terminated when the initiator quit), and the Y-axis depicts the accumulated $PR$ (Expression (3)) incurred by each user as a result of the information exchange, which also corresponds to the information each user obtained from the other. The two horizontal lines (accumulated $PR$ of the respective data sets of the two users) upper-bound the information that can be exchanged by the users.

In both settings we can see that the accumulated risk incurred by each user grows proportionally to that of his/her peer, which corresponds to a fair and balanced information exchange of sensitive data. The shape of the accumulated $PR$ curves shows that, the greater the number of iterations, the more information (and, thus, the more risk) the exchanged data involve. In the first setting (with $u_a^{HIV}$ and $u_b^{HEP}$), 43 iterations were executed before $u_a^{HIV}$ quit. The initiator quit because $u_b^{HEP}$ (whose data have a lower accumulated $PR$) was unable to provide a data piece matching the $PR$ of the last data piece sent by $u_a^{HIV}$. In the second setting (with $u_a^{HEP}$ and $u_b^{HIV}$), 47 iterations were executed. In this case, the protocol ended because $u_b^{HIV}$ (whose data have a higher accumulated $PR$) was always able to match the $PR$ of the data pieces sent by $u_a^{HEP}$; therefore, $u_a^{HEP}$ ended disclosing all her data to $u_b^{HIV}$, which explains why more iterations were executed than in the first setting. In both settings, the protocol ended once the upper bound set by the lowest accumulated $PR$ (that

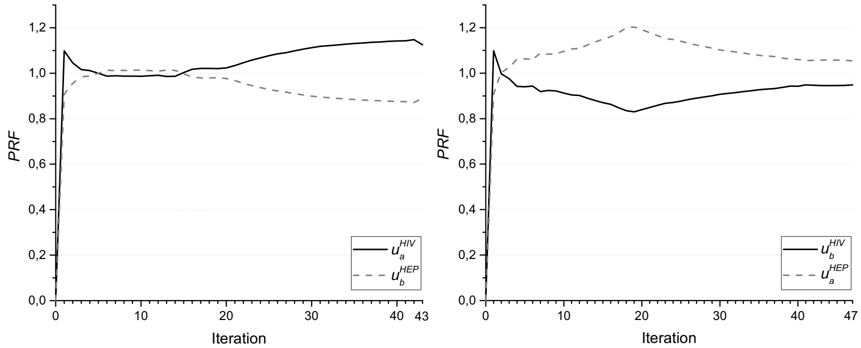

Figure 2: Protocol 1: evolution of users' $PRF$. Left: setting with $u_a^{HIV}$ and $u_b^{HEP}$. Right: setting with $u_a^{HEP}$ and $u_b^{HIV}$.

of user $u^{HEP}$) was reached. Hence, the user with the lowest accumulated $PR$ was only able to acquire as much sensitive information from his/her peer as the information contained in his/her own data. Furthermore, there was co-utility in the sense discussed in Section 4.3: to get additional data, each user had to disclose more of his/her data.

In Figure 2, we see that the $PRF$ (Expression (4)) each user derives from the execution of the protocol grows quickly and becomes stable around 1, which shows an equilibrium between the amount of disclosed and acquired information. However, both in Figures 1 and 2 we observe small differences between the $PR$ and $PRF$ of the two users (around 10-20%), which are caused by the difficulty of exchanging individual NPs that perfectly match the expected $PR$.

The second experiment was carried out with Protocol 2 (Section 4.4), which incorporates user reputations. Two different values for the responder user's initial reputation, $s_b^{ini}$, were considered: 10 (which roughly corresponds to the average information content of the NPs of each user's data) and 20 (which corresponds to the information content of the most informative NPs). Since $s_b$ measures the disclosure risk incurred by $u_b$ in previous information exchanges, it is used in Protocol 2 to compensate the potential initiator's fear of $u_b$ not reciprocating and, also, to make the information exchange more straightforward: instead of exchanging the least informative data piece at each iteration, $u_a$ can disclose an amount of data conveying as much information as $s_b$. To reinforce this latter aspect, in this experiment we allowed the users to exchange groups of NPs at each iteration. Moreover, as stated in Section 4.4, since $u_b$ is interested in maintaining his reputation high, $u_b$ will reciprocate $u_a$'s data with data pieces that best approximate $u_a$'s $PR$ *in excess*. This is the most rational behavior for $u_b$, because it increases $u_b$'s reputation and, thus, motivates $u_a$ to disclose further data faster.

Figures 3 and 4 depict the accumulated $PR$ and $PRF$ of the users in both settings according to who was the protocol initiator, $u_a$, and to the initial reputation value, $s_b^{ini}$, of the responder user $u_b$.

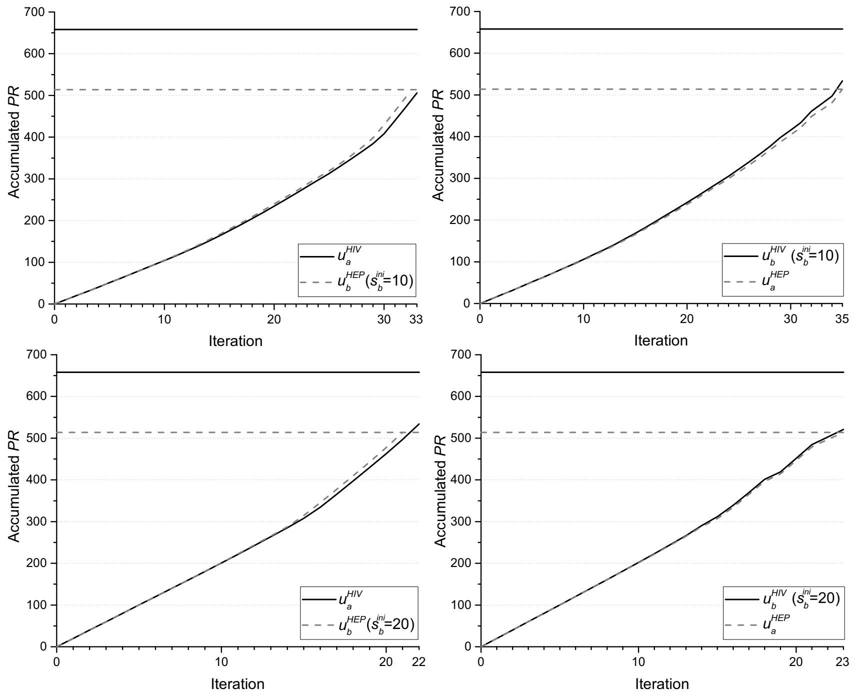

Figure 3: Protocol 2 between two users. Horizontal lines: accumulated $PR$ of the data sets held by the two users. Non-horizontal lines: accumulated $PR$ for each user as a function of the number of iterations of Protocol 2. Top left: initiator $u_a^{HIV}$, responder $u_b^{HEP}$ and responder's initial reputation $s_b^{ini} = 10$; top right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 10$; bottom left: $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 20$; bottom right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 20$.

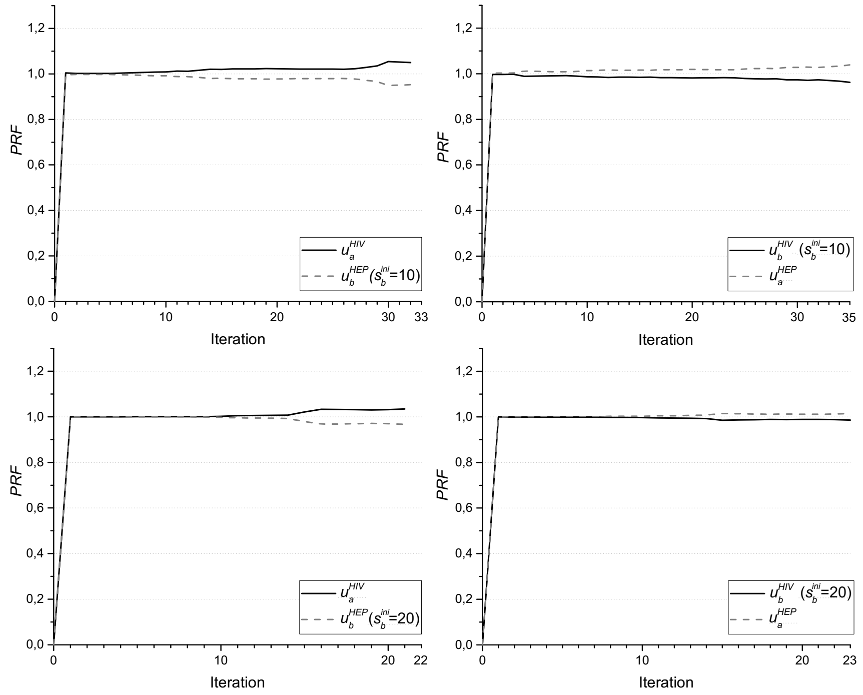

Figure 4: Protocol 2. Evolution of users' $PRF$. Top left: $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 10$; top right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 10$; bottom left: $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 20$; bottom right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 20$.

We observe two main differences w.r.t. the results reported for Protocol 1: (i) the number of iterations executed with Protocol 2 is smaller, and decreases proportionally to the initial reputation of $u_b$, and (ii) the differential between the $PR$ and $PRF$ of both users is much lower (in fact, negligible). Regarding the number of iterations, increasing the initial reputation of $u_b$ causes $u_a$ to disclose information faster than with Protocol 1, which results in less iterations and, therefore, in a more efficient information exchange. Regarding the negligible differences between the $PR$ and $PRF$ of both users, the possibility of exchanging groups of NPs enables $u_b$ to better approximate with his own data the $PR$ of the data received from $u_a$, thereby minimizing the disclosure/utility mismatch between the users in the long term.

In Figure 5 we also depict the evolution of $u_b$'s reputation in the different scenarios, which reflect the $PR$ differential incurred in past iterations. In general, $u_b$'s reputation tends to increase because $u_b$ reciprocates $u_a$'s data in excess. In the case $u_a^{HIV}$, $u_b^{HEP}$ (where the responder holds less information than the initiator), the responder's reputation increases until the responder is unable to reciprocate the initiator's data because the responder has already disclosed all his data; at this point, the responder's reputation decreases to reflect the (large) disclosure differential incurred in the last iteration of the protocol. In contrast, in the case $u_a^{HEP}$, $u_b^{HIV}$ (where the responder holds more information than the initiator), the responder's reputation keeps increasing until the end, because the responder is always able to reciprocate the initiator's data.

So far, we have considered users that follow the protocols in a rational and deterministic way. To demonstrate that our protocols favor such rational behavior and are thus co-utile (because of the mutual benefits derived by the users), we also tested Protocol 2 against the following irrational behaviors:

- *Irrationally selfish responder.* $u_b$ systematically provides his least informative data pieces to $u_a$, regardless of the $PR$ of the data received from $u_a$. In this case, $u_b$ is trying to abuse the protocol in order to unfairly increase his $PRF$ by lowering his $PR$. The results for this scenario (accumulated $PR$ and evolution of $u_b$'s reputation) are depicted in Figure 6 (taking $s_b^{ini} = 20$).

- *Random responder.* $u_b$ responds to $u_a$ by providing data pieces randomly, regardless of their $PR$ and the $PR$ of the data received from $u_a$. The results for this scenario are depicted in Figure 7 (also taking $s_{ini}^b = 20$).

When $u_b$ acts irrationally selfishly, $u_a$ quits the protocol after just two iterations. In the first one, $u_a$ sends $u_b$ data pieces involving a $PR$ around $u_b$'s high initial reputation ($s_b^{ini} = 20$), whereas $u_b$ reciprocates with a data piece whose $PR$ is much lower ($PR \approx 3$); as a consequence, $u_b$'s reputation is updated to reflect the large $PR$ mismatch. Then, in the second iteration $u_a$ sends $u_b$ a data piece involving a much lower $PR$, to which $u_b$ responds with an equally uninformative data piece. After this second iteration, $u_a$ does not have any more data pieces with low enough $PR$ and quits the protocol. In addition to $u_b$ not being able to acquire more information from $u_a$, the very low reputation

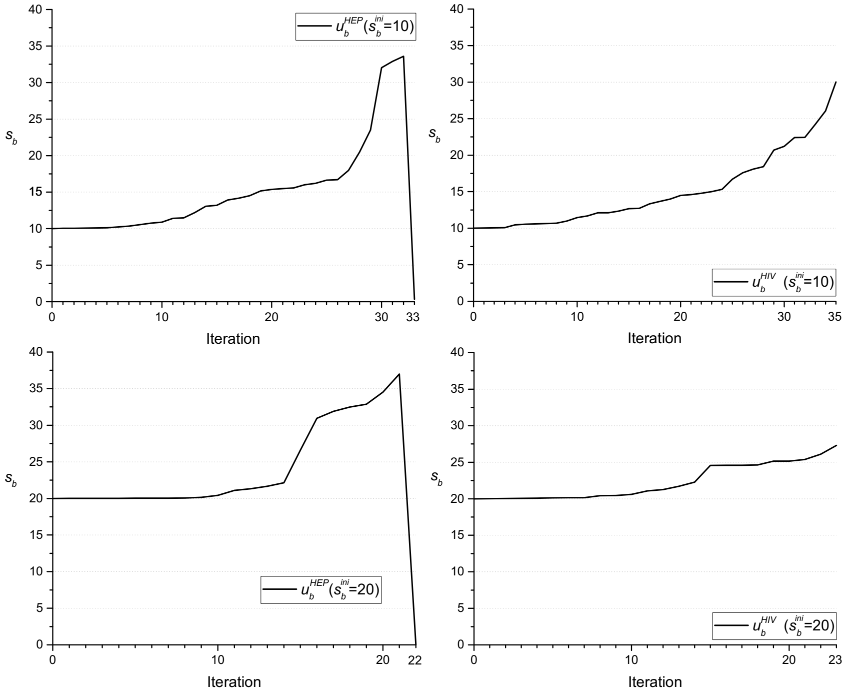

Figure 5: Protocol 2. Evolution of the reputation of the responder $u_b$. Top left: $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 10$; top right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 10$; bottom left: $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 20$; bottom right: $u_a^{HEP}$, $u_b^{HIV}$ and $s_b^{ini} = 20$.

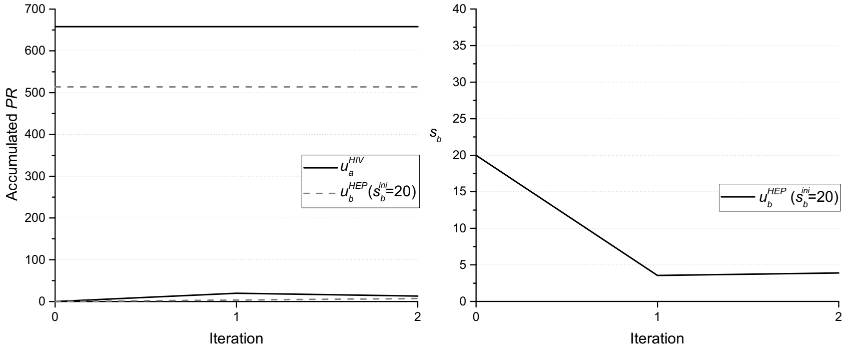

Figure 6: Protocol 2 with $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 20$, where $u_b$ behaves in an irrationally selfish way. Left graph: accumulated $PR$ for each user as a function of the number of iterations (non-horizontal lines) and accumulated $PR$ for the data sets held by each user (horizontal lines). Right graph: evolution of $u_b$'s reputation.

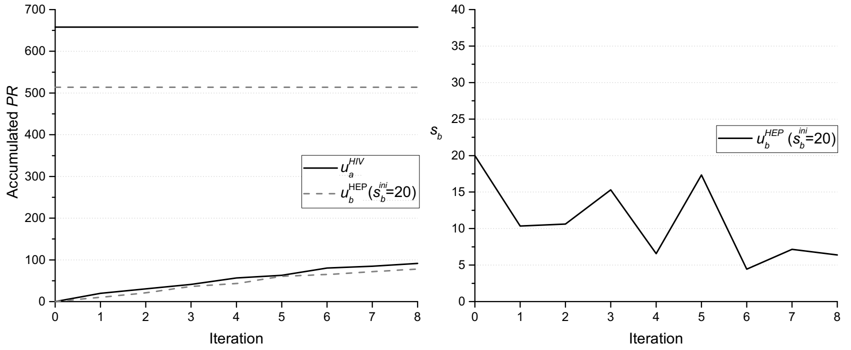

Figure 7: Protocol 2 with $u_a^{HIV}$, $u_b^{HEP}$ and $s_b^{ini} = 20$, where $u_b$ behaves in a random way. Left graph: accumulated $PR$ for each user as a function of the number of iterations (non-horizontal lines) and accumulated $PR$ for the data sets held by each user (horizontal lines). Right graph: evolution of $u_b$'s reputation.

resulting from his irrationally selfish behavior will make it very difficult for him to acquire information from other peers. Therefore, the protocol design prevents peers from rationally abusing it.

When $u_b$ acts randomly, we observe a similar behavior: $u_b$'s reputation is continuously updated according to the $PR$ mismatch resulting from each information exchange. The protocol stops when $s_b$ is lower than the $PR$ of $u_a$'s data pieces.

## 6. Protocol extensions

### *6.1. Group-based information exchange*

The protocols described so far implement a strict P2P interaction between two users. However, by following the interaction model of distributed SNs, the one-to-many information exchanges that occur in standard SNs like Facebook (in which a user publishes content and many users consume it) can be implemented as many one-to-one transactions. As discussed in Section 4.4, since our protocols can be executed automatically once the privacy requirements have been set, this iterative one-to-one exchange will not constitute an additional overhead for the user.

Even so, strict one-to-many and many-to-one information exchanges (in which the information is simultaneously and uniformly sent to a set of users) can also be supported by extending our protocols. Specifically, we assume (as is the case for real SNs like Facebook) that users are categorized in groups (*e.g.*, followers, friends, family), each one corresponding to different intimacy levels w.r.t. a concrete user, and that users must be logged in to access the information. The following changes are needed to extend our protocols to this scenario:

- In one-to-many and many-to-one information exchanges, the utility derived by user $u_a$ (former Expression (4)) from a specific group of users $u_1, \ldots, u_N$ (*e.g.*, those categorized as *friends* by $u_a$) can be rewritten as follows:

$$PRF(u_a, \{u_1, \ldots, u_N\}) = \begin{cases} \frac{\sum_{j=1, j\neq a}^{N} PR_{\mathcal{T}}(T^j)}{N \times PR_{\mathcal{T}}(T^a)} & \text{if } PR_{\mathcal{T}}(T^a) > 1, \\ \sum_{j=1, j\neq a}^{N} PR_{\mathcal{T}}(T^j) & \text{otherwise.} \end{cases} \tag{7}$$

  In plain words, the privacy risk incurred by $u_a$ grows linearly with the number $N$ of users receiving her data, whereas the information acquired by $u_a$ is the *sum* of the informativeness of all the data disclosed by the members of the group to $u_a$.

- By using Expression (7) in Protocols 1 and 2, at each iteration $u_a$ will request the users in the group to release a total amount of information that compensates $u_a$'s disclosure $N$ times. However, the information disclosed by each member of the group does need to be exactly the same.

- Even though the total information supplied by the users in the group is what matters to the initiator, reputations are individually updated in the extended Protocol 2. In this way, any users systematically disclosing significantly less informative data pieces than the average (hoping to free-ride on what the other group members disclose) will be duly penalized. For example, these users may be expelled from the group or be "degraded" to a less intimate/trusted group.

### *6.2. Compensating imbalances between users*

Up to here, we have assumed that disclosure risks can be homogeneously computed for all users in the network. Since the information exchange is reciprocal, this holds for average SN users, who are more or less "unknown" to the society in general; however, the assumption becomes less tenable for information exchanges between users having significantly different levels of social exposure. For example, disclosing that a politician suffers from a sexually transmitted disease would be of higher sensitivity than disclosing the same about an unknown citizen, because the consequences for the former may be significantly more severe than for the latter. In this case, we can say that users are "unbalanced" (because of their different social exposures) and the assessment of the privacy risks they incur when releasing information should reflect the imbalance.

The exposure of a user can be regarded as being proportional to her social influence; in the context of SNs, the influence of a user can be roughly quantified by the number of users *following* her. Thus, we propose to use the *number of followers* to weight the privacy risks the users incur when disclosing data; in this way, the risk assessment can account for the different social exposure levels of the interacting users.

Let followers($u_a$) and followers($u_b$) be the number of followers of $u_a$ and $u_b$ in the SN, respectively. The *normalized privacy risk* ($NPR$) the users incur when exchanging information is balanced by weighting Expression (2) by their number of followers. Specifically, for $u_a$ the $NPR$ is computed as follows:

$$NPR_{\tau}(t^a) = \frac{\text{followers}(u_a)}{\max(\text{followers}(u_a), \text{followers}(u_b))} \times PR_{\tau}(t^a); \tag{8}$$

for $u_b$ it is as follows:

$$NPR_{\tau}(t^b) = \frac{\text{followers}(u_b)}{\max(\text{followers}(u_a), \text{followers}(u_b))} \times PR_{\tau}(t^b). \tag{9}$$

Expressions (8) and (9) can be directly used instead of Expression (2) in Protocols 1 and 2, in order to weight the assessment of disclosure risks between peers with significantly different levels of social exposure.

## 7. Conclusions and future work directions

The privacy concerns of users are a major threat to the sustainability of SNs, especially of those with very sensitive scopes. In this paper, we have formalized the utility that rational users derive from participating in SNs. Also, we have argued that the current information exchange model, in which users release (sensitive) information to others in an uncontrolled way, contradicts the rational interests of privacy-aware users. To solve this issue, and thereby ensure the sustainability of SN in the long term, we can leverage decentralized SNs (in which users are not concerned anymore about the SN operator getting hold of their personal data) and the P2P *co-utile* information exchange protocols we propose. Our protocols ensure that the exchange of information is self-enforcing and mutually beneficial even for privacy-aware users. Our empirical work confirms that our protocols favor rational and mutually beneficial behaviors, and are immune to rational attempts to abuse them. Moreover, we have also shown that incorporating user reputations (managed in an also decentralized and co-utile way) contributes to mitigating the reluctance of users to disclose sensitive information to other peers and makes the information exchange more efficient.

As future work, we plan to develop protocols for detecting malicious behaviors and punish users through their reputations, as we discuss in Section 4.4; in this way, users will be rationally motivated to avoid malicious actions, which will ensure that co-utility holds. On a technical side, we plan to develop plugins for well-known SNs that, by relying on our automatic disclosure assessment, educate users on the privacy risks inherent to data release. We also plan to implement our protocols in decentralized SNs, such as Diaspora, in order to validate their suitability and scalability, and also to obtain feedback from real users and daily use. Last but not least, we intend to investigate other utility and privacy definitions: whereas the information acquired on other users is a reasonable definition of utility in some SNs (for example, professional or health-care SNs), the utility of the typical Facebook user may be also related to social visibility [25]. Work is needed to formalize an entire range of plausible utility functions, taking into account that changing the definition of utility may also have an impact on the definition of privacy.

### Acknowledgments and disclaimer

Funding by the Templeton World Charity Foundation (grant TWCF0095/AB60 “CO-UTILITY”) is gratefully acknowledged. Also, partial support to this work has been received from the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer), the Spanish Government (projects TIN2014-57364-C2-1-R “SmartGlacis”, TIN2015-70054-REDC and TIN2016-80250-R “Sec-MCloud”) and the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”). The authors are with the UNESCO Chair in Data Privacy, but the views in this work are the authors’ own and are not necessarily shared by UNESCO or any of the funding bodies.

### References

[1] Abril, D., Navarro-Arribas, G., Torra, V. (2011) On the declassification of confidential documents, in Modeling Decisions for Artificial Intelligence – MDAI 2011, LNCS 6820, pp. 235-246, Springer.

[2] Anandan, B., Clifton, C. (2011) Significance of term relationships on anonymization, in IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, pp. 253-256, Lyon, France.

[3] Babaioff, M., Chuang, J., Feldman, M. (2007) Incentives in peer-to-peer systems, in N. Nisan, T. Roughgarden, É. Tardos and V. V. Vazirani (eds.), Algorithmic Game Theory, pp. 593-611, Cambridge University Press.

[4] Batet, M., Sánchez, D. (2016) Improving semantic relatedness assessments: ontologies meet textual corpora, in Proceedings of the 20th International Conference KES 2016, pp. 365-374.

[5] Becker, J., Chen, H. (2009) Measuring privacy risk in online social networks, in Web 2.0 Security and Privacy Conference.

[6] Buccafurri, F., Coppolino, L., D’Antonio, S., Garofalo, A., Lax, G., Nocera, A., Romano, L. (2014) Trust-based intrusion tolerant routing in wireless sensor networks, in Proceedings of the International Conference on Computer Safety, Reliability and Security SAFECOMP 2014, pp. 214-229.

[7] Buccafurri, F., Fotia, L., Lax, G., Saraswat, V. (2016) Analysis-preserving protection of user privacy against information leakage of social-network Likes, Information Sciences, 328:340-358.

[8] Carminati, B., Ferrari, E., Heatherly, R., Kantarcioglu, M., Thuraisingham, B. (2011) Semantic web-based social network access control, Computers and Security, 30:108-115.

[9] Carminati, B., Ferrari E., Perego, A. (2007) Private relationships in social networks, in Proceedings of ICDE'07 Third International Workshop on Privacy Data Management, pp. 163-171, IEEE Computer Society.

[10] Cheek, G.P., Shehab, M. (2012) Privacy management for online social networks, in 21st International Conference Companion on World Wide Web – WWW'12, pp. 475-476.

[11] Chow, R., Golle, P., Staddon, J. (2008) Detecting privacy leaks using corpus-based association rules, in 14th Conference on Knowledge Discovery and Data Mining, pp. 893-901, Las Vegas, NV: ACM.

[12] Church, K.W., Hanks, P. (1990) Word association norms, mutual information, and lexicography, Computational Linguistics, 16:22-29.

[13] Consumer Reports National Research Center (2010) State of the net 2010, Consumer Reports Magazine, June 2010.

[14] Cutillo,, L., Molva, R., Strufe, T. (2009) Safebook: a privacy-preserving online social network leveraging on real-life trust, IEEE Communications Magazine, 47:94101.

[15] D'Arcy, J. (2011) Combating cyber bullying and technologys downside, The Washington Post, Sep. 21.

[16] Daud, M.I., Sánchez, D., Viejo, A. (2016) Privacy-driven access control in social networks by means of automatic semantic annotation, Computer Communications, 76:12-25.

[17] Dhia, I., Abdessalem, T., Sozio, M. (2012) Primates: a privacy management system for social networks, in 21st ACM international conference on Information and knowledge management – CIKM'12, pp. 2746-2748.

[18] Domingo-Ferrer, J. (2010) Rational privacy disclosure in social networks, in Modeling Decisions for Artificial Intelligence – MDAI 2010, LNCS 6408, pp. 255-265. Springer.

[19] Domingo-Ferrer, J., Farràs, O., Martínez, S., Sánchez, D., Soria-Comas, J. (2016) Self-enforcing protocols via co-utile reputation management, Information Sciences, 367-368: 159-175.

[20] Domingo-Ferrer, J., Martínez, S., Sánchez, D., Soria-Comas, J. (2017) Co-utility: self-enforcing protocols for the mutual benefit of participants, Engineering Applications of Artificial Intelligence, 59:148-158.

[21] Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J. (2016) Co-utility: self-enforcing collaborative protocols with mutual help, Progress in Artificial Intelligence, 5(2):105-110.

[22] Domingo-Ferrer, J., Viejo, A., Sebé, F., González-Nicolás, Ú. (2008) Privacy homomorphisms for social networks with private relationships, Computer Networks, 52:3007-3016.

[23] The European Parliament and the Council of the EU (1995) Data Protection Directive 95/46/EC.

[24] Gerlach, J., Widjaja, T., Buxmann, P. (2015) Handle with care: How online social network providers privacy policies impact users information sharing behavior, The Journal of Strategic Information Systems, 24(1):33-43.

[25] Krasnova, H., Spiekermann, S., Koroleva, K., Hildebrand, T. (2010) Online social networks: why we disclose, Journal of Information Technology, 25(2):109-125.

[26] Lee, B., Fan, W., Squicciarini, A.C., Ge, S., Huang, Y. (2014) The relativity of privacy preservation based on social tagging, Information Sciences 288:87-107.

[27] Lemaire, B., Denhière, G. (2006) Effects of high-order co-occurrences on word semantic similarities, Current Psychology Letters - Behaviour, Brain and Cognition, 18(1):1.

[28] Liu, K., Terzi, E. (2009) A framework for computing the privacy scores of users in online social networks, in Proceedings of ICDM 2009-The 9th IEEE International Conference on Data Mining, pp. 288-297.

[29] Narayanan, A., Shmatikov, V. (2009) De-anonymizing social networks, in 30th Symposium on Security and Privacy, pp. 173-187, Oakland CA.

[30] Nilizadeh, S., Jahid, S., Mittal, P, Borisov, N., Kapadia A. (2012) Cachet: a decentralized architecture for privacy preserving social networking with caching, in 8th International Conference on Emerging Networking Experiments and Technologies – CoNEXT 2012, pp. 337-348, ACM.

[31] Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy, in 14th International Joint Conference on Artificial Intelligence–IJCAI 1995, pp. 448-453, Morgan Kaufmann Publishers Inc.

[32] Sánchez, D., Batet, M. (2011) Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective, Journal of Biomedical Informatics, 44(5):749-759.

[33] Sánchez, D., Batet, M. (2016) C-sanitized: A privacy model for document redaction and sanitization, Journal of the Association for Information Science and Technology, 67(1):148-163.

[34] Sánchez, D., Batet, M., Viejo, A. (2013) Minimizing the disclosure risk of semantic correlations in document sanitization, Information Sciences, 249: 110-123.

[35] Sánchez, D., Batet, M., Viejo, A. (2013) Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security, 8:853-862.

[36] Sánchez, D., Batet, M., Viejo, A. (2014) Utility-preserving privacy protection of textual healthcare documents, Journal of Biomedical Informatics, 52: 189-198.

[37] Sánchez, D., Viejo, A. (2015) Privacy risk assessment of textual publications in social networks, in International Conference on Agents and Artificial Intelligence - ICAART'15, pp. 236-241.

[38] Snowdon, G. (2011) The rules of social recruiting, The Guardian, Aug. 19.

[39] Srivastava, A., Geethakumari, G. (2013) Measuring privacy leaks in online social networks, in 2013 International Conference on Advances in Computing, Communications and Informatics-ICACCI, IEEE.

[40] Staddon, J., Huffaker, D., Larking, B., Sedley, A. (2012) Are privacy concerns a turn-off?: engagement and privacy in social networks, in 8th Symposium on Usable Privacy and Security, article no 10. ACM.

[41] Stern, T., Kumar, N. (2014) Improving privacy settings control in online social networks with a wheel interface, Journal of the Association for Information Science and Technology, 65:524-538.

[42] Talukder, N., Ouzzani, M., Elmagarmid, A.K., Elmeleegy, H., Yakout, M. (2010) Privometer: privacy protection in social networks, in IEEE 26th International Conference on Data Engineering Workshops-ICDEW, IEEE.

[43] Tardos, É., Vazirani, V.V. (2007) Basic solution concepts and computational issues, in N. Nisan, T. Roughgarden, É. Tardos and V. V. Vazirani (eds.), Algorithmic Game Theory, pp. 3-28, Cambridge University Press.

[44] Terry, N., Francis, L. (2007) Ensuring the privacy and confidentiality of electronic health records, University of Illinois Law Review, pp. 681-735.

[45] Trujillo-Rasua, R., Yero, I.G. (2016) $k$-Metric antidimension: a privacy measure for social graphs, Information Sciences 328:403-417.

[46] U.S. Federal Trade Commission (2014) Data Brokers. A Call for Transparency and Accountability.

[47] Vaughan-Nichols, S. (2010) Diaspora: Its no facebook ... yet, Computerworld, Sep. 16.

[48] Viejo, A., Sánchez, D. (2016) Enforcing transparent access to private content in social networks by means of automatic sanitization, Expert Systems with Applications, 62:148-160.

[49] Wang, Y., Nepali, R.K. Nikolai J. (2014) Social network privacy measurement and simulation, in International Conference on Computing, Networking and Communications – ICNC'14, pp. 802-806, Honolulu, HI: IEEE Computer Society.

[50] Zhang, C., Sun, J., Zhu, X., Fang, Y. (2010) Privacy and security for online social networks: challenges and opportunities, IEEE Network, 24:13-18.