Chapter 14
Protecting Privacy in Volunteered Geographic Information Processing

Marc Löchner, Alexander Dunkel, and Dirk Burghardt

Abstract Social media data is used for analytics, e.g., in science, authorities, or the industry. Privacy is often considered a secondary problem. However, protecting the privacy of social media users is demanded by laws and ethics. In order to prevent subsequent abuse, theft, or public exposure of collected datasets, privacy-aware data processing is crucial. In this chapter, we show a set of concepts to process social media data with social media user's privacy in mind. We present a data storage concept based on the cardinality estimator HyperLogLog to store social media data, so that it is not possible to extract individual items from it, but only to estimate the cardinality of items within a certain set, plus running set operations over multiple sets to extend analytical ranges. Applying this method requires to define the scope of the result before even gathering the data. This prevents the data from being misused for other purposes at a later point in time and thus follows the privacy by design principles. We further show methods to increase privacy through the implementation of abstraction layers. As another additional instrument, we introduce a method to implement filter lists on the incoming data stream. A conclusive case study demonstrates our methods to be protected against adversarial actors.

Keywords Privacy · Social media · Data retention · HyperLogLog

14.1 Introduction

Social media services like Twitter or Instagram are used to communicate and share information worldwide, which generates a rich set of data. Since a large part of this data is publicly available, it can be used beyond the features of the social media services itself, especially by third parties.

M. Löchner (✉) · A. Dunkel · D. Burghardt
Technische Universität Dresden, Dresden, Germany
e-mail: marc.loechner@tu-dresden.de; alexander.dunkel@tu-dresden.de; dirk.burghardt@tu-dresden.de

D. Burghardt et al. (eds.), *Volunteered Geographic Information*,
https://doi.org/10.1007/978-3-031-35374-1_14

The main problem with social media data utilization for applications other than their dedicated use case is that explicit consent from the social media user is usually missing. While most users are aware that their content is publicly available on the Internet, they do not assume that data is frequently recycled for other purposes such as scientific, commercial, or administrative use (Boyd and Crawford 2012). Accordingly, this demands an exceptional strong focus on their privacy.

In contrast to other environments, the data to be protected is already public (Williams et al. 2017). In the view of third parties, that data can be utilized for any purpose, including those that oppose the user's interest (Zhou et al. 2008). But data can also be used with good intentions (Daly et al. 2019), whereas "good" could be defined by "in the user's Interest." For example, social media has shown a valuable source of information in crisis mapping, emergency response, or public planning (Fiedrich and Fathi 2021; Dunkel 2021).

In order to support ongoing development of positive use cases, scientists need to respect and actively protect social media users' privacy. Scientists need to take explicit control over data that they expose and prevent accidental disclosures.

An approach to support the adoption of accidental disclosure prevention techniques is to *prevent* the gathering of privacy-relevant data in the first place. We specifically aim at providing methods for the use of social media data following the *privacy-by-design* principles (Cavoukian et al. 2009).

In this chapter, we show a set of concepts that enable to process social media data with social media user's privacy in mind. We present a data storage concept that implements an algorithm called HyperLogLog (HLL) (Flajolet et al. 2007) to not store raw social media data but only statistics about their occurrence. We further show that while losing precision of the data, privacy can even be increased by applying multiple layers of abstraction on the data. For a context-dependent treatment of privacy and to cover edge cases, we further introduce a model to implement filter lists on the incoming data stream. A conclusive case study demonstrates our methods to be protected against adversarial actors.

14.2 Fundamentals

14.2.1 Related Work

Issues and challenges related to privacy arise everywhere, where social media data is involved. Following up, we link to research projects within this book, which are primarily based on processing social media data and therefore our research is relevant for.

The EVA-VGI project (see Chap. 12) studies the heterogeneity, quality, subjectivity, spatial resolution, and temporal relevance of geo-referenced social media data. Focusing on the integration of spatial, temporal, topical, and social dimensions combined with an explicit link between events and reactions, they present concep-

tual approaches and methods that enable a privacy-aware visual analysis of VGI in general and geo-social media data in particular. The project has taken advantage of the results of our research by implementing HLL on datasets related to their publications (Dunkel et al. 2020).

Similarly, the VA4VGI project (see Chap. 6) describes how geo-aware filtering and anomaly detection on geo-referenced social media data can be a significant information source for stakeholders in journalism, urban planning, or disaster management. They present tag maps that provide overview-first, details-on-demand, visual summaries of large amounts of social media data over time and thus visualize their temporal evolution.

Closely related to the former is the DVCHA project (see Chap. 13). The overall objective of their research is to study the implications of social media data for the efficiency of disaster management. Focusing on so-called Virtual Operations Support Teams (VOST), their research addresses motivation, success factors, and improvement of distributed decision making processes based on disaster-related real-time social media data.

In a collaboration with the DVCHA project, we carried out a case study, in which we explored the deployment of HLL into disaster management processes (see Sect. 13.6). We developed and conducted a focus group discussion with VOST members, where we identified challenges and opportunities of working with HLL and compared the process with conventional techniques (Löchner et al. 2020). Findings showed that deploying HLL in the data acquisition process of VOST operations will not distract their data analysis process. Instead, several benefits, such as improved working with huge datasets, may contribute to a more widespread use and adoption of the presented technique, which provides a basis for a better integration of privacy considerations in disaster management.

14.2.2 On Privacy Aspects

From a generic point of view, *privacy* is the freedom to fully or partially retreat oneself in a self-controlled manner. There are always multiple forms of definitions of the term *privacy*, stretching from personal to a cultural point of views (Solove 2008). It is important to distinguish between the *right to privacy* and the *concept of privacy* (Hildebrandt 2006). The *right* is clearly formed by laws, whereas the *concept* is rather vaguely determined based on subjectively perceived personal values. Privacy is often sacrificed voluntarily in exchange for perceived benefits and sometimes violated by others, either intentionally or accidentally (Reyman 2013).

Privacy by design as a set of principles is a relevant objective in the conception of applications in general. As Cavoukian et al. (2009) state, privacy must be approached from a design-thinking perspective. It must be incorporated in technologies not as an optional on-top feature but as a fundamental characteristic of organizational priorities, project objectives, design processes, and planning

operations. Concepts built upon these principles are hard to break in terms of privacy violations.

A contemporary method to protect data has been presented as *differential privacy* (DP) (Dwork 2008) and adopted frequently (Desfontaines and Pejó 2020). DP adds certain amounts of random data to a set of real data set, in order to make real data indistinguishable from the random data and thus protect it from being identified as such. However, DP still requires the original data to be available to process. Furthermore, DP requires developing new concepts and models for each data set, which is very inefficient when dealing with really large sets of data.

In the geo-community, there is a wide range of concepts known to protect privacy in terms of location data. Some techniques are based on anonymity, e.g., *mix zones* (Beresford and Stajano 2003) or *k-anonymity* (Ciriani et al. 2007). Others are based on obfuscation, e.g., *imprecision* (Duckham and Kulik 2005), or policy like *restriction* (Hauser and Kabatnik 2001). All of these approaches require the possession of original raw data. Processed data sets are unable to be updated with subsequent data, which requires reprocessing of the entire data set upon updates. This is very inefficient when dealing with large amounts of social media data.

In the context of social media data, the consideration of privacy, ethics, and legal issues should play an important role. The statement "Privacy of user data and information should be considered in the initial design of VGI systems" (Mooney et al. 2017) can be extended to platforms and methods for the analysis and further processing of social media data in general.

Kounadi et al. (2018) discuss privacy threats related to inference attacks on *geosocial network data*. They provide protection recommendations for sharing these sorts of data and publishing resulting visualizations. Keßler and McKenzie (2018) proposed in a total of 21 theses to reflect on the current state of *geoprivacy* from a technological, ethical, legal, and educational perspective. They provide various examples of how common it has become to share location and how it can be used and misused.

14.2.3 Data Retention

Processing social media data is to a relevant extent based on operating analytics software, which provides automatic analysis on gathered social media data stored in local databases. Their user interfaces take input to be crawled for in the stored data and return, for example, statistics of post occurrences in any context. Depending on the situation, only parts of that information may be relevant (see Sect. 14.3.1). Still, the entirety of every post has been and remains stored in local databases.

This means that if a data item is being deleted on the site of the corresponding social media service, it still resides at the place where it has been downloaded to. Technically, that practice meets the requirements to be termed *data retention*. We define this term as such: preserving data for an indefinite time period with no specific

purpose for any individual data item but with the assumption to make use of the information in entirety at a later point in time.

The term is being discussed in the public mostly in conjunction with telecommunication analysis and surveillance. European Digital Rights public interest group states that "data retention practices interfere with the right to privacy at two levels: at the level of retention of data, and at the level of subsequent access to that data by law enforcement" (Rucz and Kloosterboer 2020).

We introduce the term in a broader and more technical environment to emphasize the explosive nature of recklessly dealing with personal data, which social media data is (European Commission 2018). According to the above definition, the term is valid for any case of storing and retending personal data in stocks. Wright et al. (2020) use it even to describe any storage of data underlying scientific studies.

Owning a set of data requires great responsibility in terms of data security. It opens up risks of possible abuse, theft, or accidental public exposure (Miller 2020). Breaking it down to a simple rule, it can be stated that "the more data you have, the more data you can lose" (Guillou and Portner 2020).

Beyond governmental agencies and law enforcement, also commercial players, journalists, researchers, or nonprofit organizations face challenges when storing individual-related data like those from social media. Stieglitz et al. (2018) discovered that the volume of data was most often cited as a challenge by researchers. Wang and Ye (2018) summarize common techniques for social media analytics in natural disaster management and coin the term *mining* for that matter.

Furthermore, the social impact of misusing large sets of data is well-known. The Cambridge Analytica scandal is one of the examples that show how massive data sets can be alienated (Berghel 2018). The company used personal information from millions of Facebook users without their consent to derive information about their political points of view and then microtarget personally tailored political advertisements to them. They claimed to have a major impact on the 2016 US presidential election, which can be regarded as a threat to democratic legitimacy (Dowling 2022).

Users of social media services start to realize that all of their data is not only publicly available but made use of by third parties. Data retention drives forgetfulness as a social concept at risk (Blanchette and Johnson 2002). The *chilling effect*, people slowly increasing self-discipline and restriction of their communication behavior due to becoming aware of digital surveillance, and panopticism (Manokha 2018; Büchi et al. 2022) are described consequences.

Nevertheless, the huge amount of data raised by social media services being a tremendous privacy thread is only one side of the coin. Large sets of social media data can also be beneficial for the public. The work of humanitarian organizations depends on publicly available data that is authentic and relevant. Especially, VOSTs rely on the availability of public social media data (Kuner and Marelli 2020); therefore, its prosperity must be preserved. A gradual retreat of users from social media services in favor of closed, "antisocial" messaging groups (Leetaru 2019; Wilson 2020) must be prevented.

14.2.4 HyperLogLog

One of our contributions to this issue presented in this chapter is based on storing data using an algorithm called *HyperLogLog* (HLL). This algorithm is a cardinality estimator first introduced by Flajolet et al. (2007).

Its fundamental strength is the ability to *estimate* the distinct count of a multiset (cardinality) and store it in a data structure, which does not allow the extraction of individual elements. This is done by storing only hashes of data items instead of the original raw data and identifying them by counting leading zeros of the binary representation of their hashes. The algorithm is able to predict how many distinct items have been added to the HLL set, based on the maximum number of leading zeros observed. This makes processing data using HLL very efficient in terms of processing time and storage space. It is not possible to search for prior unknown information in an HLL set, for example, the usernames of all the posts that have been gathered. This makes implementing HLL follow the *privacy by design* principle.

14.3 Concepts

14.3.1 Privacy-Aware Storage

The key aspect for our approach is to make it impossible to relate to the original social media data from a given processed data set (*privacy by design*). Therefore, we propose to utilize the cardinality estimation algorithm HyperLogLog (HLL) described in Sect. 14.2.4 to gathered store social media data.

To provide a minimal example of the process, we introduce a scenario, in which the difference in spatial occurrences of social media posts including a certain hashtag should be visualized. The result should be a choropleth map of areas according to the amount of post occurrences within that area (see Fig. 14.1). Areas are defined by a *GeoHash*, a hierarchical grid-like geocode identification concept (Niemeyer 2008; Morton 1966).

To store the occurrence of posts in an area, it is only necessary to *count* the number of distinct occurring posts, their *cardinality*. Reflecting, this unveils that storing the entirety of a social media post is unnecessary. It is sufficient to memorize its unique identifier (ID), which has been assigned by the social media service it originates from.

However, storing the ID in clear text in the database will allow identifying the post and thus the author of a post later on. The characteristics of HLL in turn enable to store data like the ID of a post in a set without the ability to regain it without prior knowledge about its existence in the set. Storing post IDs in an HLL set related to their geohash will only reveal their cardinality. Posts that occur later in the stream and match the same geohash will be added to this HLL set, which increases its cardinality by one for each new post. The geohash itself representing the post's

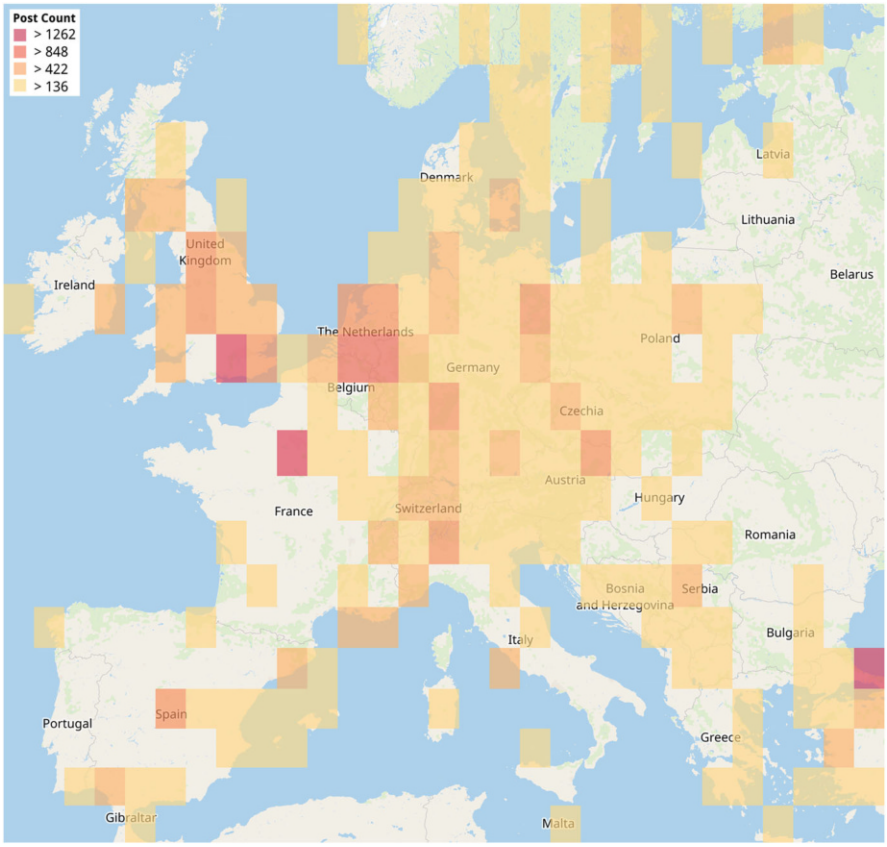

Fig. 14.1 Example of a map showing areas with different occurrences of posts containing `#omicron` hashtag on Twitter from January through March 2022. Map data: OpenStreetMap contributors Color distribution: Head/Tail Breaks (Jiang 2013)

Table 14.1 Exemplary database table structure showing four records (each stands for one area represented by the geohash) and the corresponding HLL set containing the post IDs

geohash	id
w41s	\x128b7fdf939b45ec2ef0ca
6yws	\x128b7fbfd17eca803517d2
c29s	\x128b7fe00ef312fcf023c9
75cs	\x128b7fcc47a6c00361c5e7

originating area is stored as the index of the database record (see Table 14.1). The resulting HLL data structure represents all posts matching a certain term from a certain area, while it is impossible to derive the post IDs back from it.

Using HLL, we do *not* store the post IDs itself but calculate hashes from them and store them in an array of counters that represent the set of post IDs (see

Sect. 14.2.4). Table 14.1 shows an example database table structure with geohash values representing an area and the corresponding HLL set representing the IDs of posts that occurred in that area.

Having a database with geohashes and their corresponding HLL set as shown exemplarily in Table 14.1, it is possible to compute the cardinality of the HLL set and thus determine the number of posts in each area. The result of such a computation could as well be achieved by just incrementing an integer per seen post ID and storing the sum instead of an HLL set. The significance of using the HLL algorithm instead is that it provides the opportunity to perform the set operations *union* and *intersection* on the HLL sets.

This can be useful for combinations of individual data sets. Different sets of gathered posts, each relating to certain terms, can be combined to monitor a more specific scenario.

A social media post as a data item can be broken down into its spatial, temporal, topical, and social components, each of which can be stored as separate HLL sets. As shown in Fig. 14.2, this can lead to a number of different HLL sets, each containing the post IDs of posts matching different criteria: involving a certain topic, originating in a certain area or in a certain time period, or authored by a user of a certain group.

Using the topical facet exemplarily in a disaster management scenario, an intersection of a set containing posts with the terms `fire` and one containing `forest` posts could lead more precisely to disaster incidents than both terms on their own. It still makes sense to monitor the terms individually in the first place because a combination of `fire` and `accident` can lead to other and different disaster incidents, as well as `forest` and `accident` does.

Furthermore, different terms could have the same meaning, for example, `flood`, `high tide`, `wave`, and `tsunami` could all refer to the same situation. So, a union of HLL sets on posts over these terms can provide more comprehensive information about disasters. Likewise, terms in different languages could also be

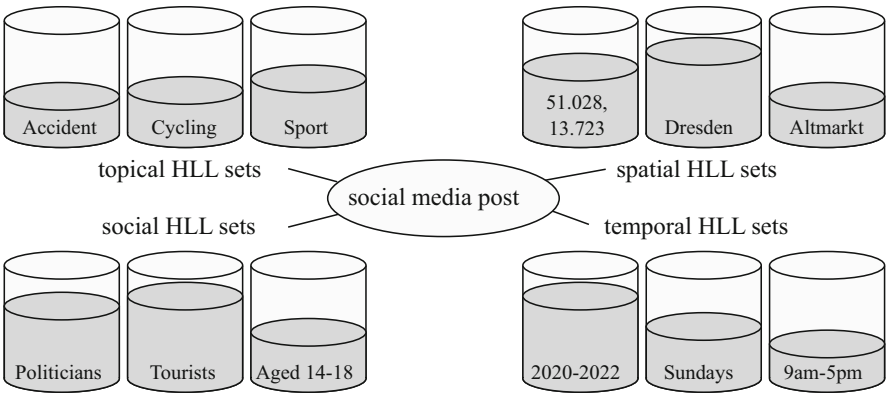

Fig. 14.2 Examples of HLL sets derived from the four facets of a social media post

monitored in combination. This, for example, enables VOSTs (see Sect. 14.2.1 and Chap. 13) to monitor larger, multiple languages involving areas like border triangles or including smaller countries like Benelux or the Baltics.

This concept provides privacy by design because it does not store the post IDs in a readable way. It only stores a statistical derivative resulting from the characteristics of the HLL algorithm (see Sect. 14.2.4) and complies to the *privacy by design* principles. The following subsections cover how it can be extended even further by applying extended concepts to adjust the level of privacy protection.

14.3.2 Abstraction Layers

The concept of *abstraction* has been widely used in the geo-community to visualize spatial information scale dependent on different degrees of detail (Burghardt et al. 2016). We re-dedicate these generalization methods from geovisualization to privacy protection.

Herein, we present a model to improve privacy for social media users, in particular in the context of data collection. It aims at withdrawing precision from the data by deriving multiple abstraction layers of it. Applying these layers, we are able to quantitatively describe different levels of privacy. By deploying methods of *generalization* and thus decreasing precision of the data, we can increase privacy, and vice versa.

Figure 14.3 shows a visual representation of this model, following the four-facet representation to characterize a social media post, introduced by Dunkel et al. (2019). The bottom layer in each facet is formed by the original data. Each following layer represents an increase in privacy protection for the user. This way, we have the ability to adjust the level of detail of the data in a fine-grained and context-dependent way. Each of the layers is described in detail in the following subsections.

14.3.2.1 Spatial Facet

In the spatial facet, the original data is usually represented by a *coordinate* in latitude and longitude or a tiny area surrounding that coordinate. A first abstraction from it

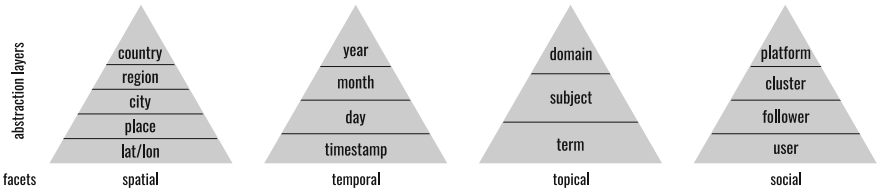

Fig. 14.3 Abstraction layers for each facet

can be an arbitrarily named *place* that includes the coordinate, e.g., a market square or a park. The next abstraction could be an administrative or functional *region* or other territories enclosing the place, e.g., a city or metropolitan area, county, or state. Cities can also be regarded as intermediate layers below regions. The next abstraction layer could be a *country* or another even broader defined, e.g., natural, political, administrative, or religious region.

When applying this model to the HLL-based storage concept presented in Sect. 14.3.1, it is crucial to note that even the lowest layer needs to be an area to be able to count multiple posts within it. If using a point coordinate instead, chances tend toward zero for multiple posts hitting that exact coordinate. An alternative approach with clustering techniques would be necessary there.

In an application implementing this model, the database index would be the geohash, place, city, or country, depending on the layer of abstraction. The corresponding HLL sets include the hashed post IDs of posts that originate in that respective area. An appropriate visualization of that data would be a map showing visually differentiable areas (see, e.g., Fig. 14.1).

14.3.2.2 Temporal Facet

Abstractions in the temporal facet are clearly defined by common time units. The basic layer is the timestamp of the publication of a post, abstracted as the day of the publication or a month or even a year.

In analogy to the spatial facet, when implementing this model, the database index must be a time range rather than a point in time, to be able to associate multiple matching posts with it. To visualize just the temporal facet, a timeline is the preferred graphical representation. Usually, this facet requires to also filter for a certain topic first, to prevent just visualizing *all* social media posts occurring within a certain time period.

14.3.2.3 Topical Facet

The topical facet is characterized by applying topic modeling techniques (Kherwa and Bansal 2019) to find abstractions of terms. The basic layer can be defined by the *terms*, the original content in the post, e.g., `The River Elbe has burst its banks in Dresden today`. A more general layer can be rendered in the overall *subject* of the post, e.g., `Dresden flood`. Another abstraction layer can be the *domain* of the `Natural Disaster`.

In an implementation, the database index will represent these terms, subjects, or domains, and the corresponding HLL sets hold the associated post IDs. It is not trivial to generate more generic terms for specific posts, but topic modeling techniques can help with that task. A word cloud can visualize these terms in different sizes (Hearst et al. 2019) depending on the cardinality of posts associated with them.

14.3.2.4 Social Facet

The social facet relates to users of social media. When running analyses on this object, it is crucial to note that we are switching the focus. While we are trying to avoid storing personal data around an object in the other facets, here we want to achieve the opposite: counting appearances of data that relates to a single person or a group. An exemplary analysis would be to count the number of posts per user or group. In this scenario, it is especially useful to apply abstraction layers in order to gain privacy for a single user. In analogy to the temporal facet, it is also useful to filter posts for a certain topic in beforehand.

In the basic layer, the creator of a social media post, the *user*, is targeted. The database index would be the username or id, and the corresponding HLL set consists of the post IDs. Combining, e.g., all of the user account's *followers* to a group and regarding posts of all of them could apply as the first abstracted layer. Through network analysis (Maireder et al. 2014), we can define *clusters* to be objects in a second step of abstraction. Another layer of abstraction could be the consideration of different social network *platforms* (Cosenza 2022), which have a distinct user base, which might originate from different cultural backgrounds, e.g., Twitter, Instagram, WeChat, and VKontakte.

Implementing groups of users is more challenging than in other layers. The group of followers in the second layers consists of a list of user IDs eventually, which could again be stored in an HLL set and get an ID assigned to. IDs of multiple groups are then stored in HLL sets and can be combined or contrasted with other group IDs, defining clusters accordingly.

All the described layers are only examples and can be replaced by other structures. Also, the number of abstraction layers can be chosen arbitrarily, as the granularity of the data can change.

It should be noted that abstraction layers do not only gain privacy for the social media users, but they also diminish the precision of the data. This makes applying abstraction to social media data be a compromise between privacy and precision.

14.3.3 Filter Lists

Storing social media data using HLL to be processed in analytics software forms the basement of privacy protection. Applying generalization methods as described in Sect. 14.3.2 provides further opportunities to adjust data precision. However, there are edge cases that require special handling. For instance, even the existence of a single specific term, a specific time, location, etc. may provide hints that can be repurposed or combined with other (e.g., external) information to compromise user privacy in certain situations. Following the principle that different data must be treated differently (Almås et al. 2018), we seek to contribute to a systematic approach to fine-tuning privacy preservation and analytical flexibility.

There are two main approaches to adjusting privacy—utility trade-offs with HLL and abstraction layers. First, *stop and allow lists* can be used during the generation of the HLL set to enable context-dependent data protection through filtering. Second, *threshold values* can be defined flexible to influence the granularity of the HLL set indexes and, based on that, the degree of anonymity. Table 14.2 lists examples for each context in the framework, where accuracy (utility) may be traded in favor of a higher degree of privacy, similar to the broader data sensitivity spectrum proposed by Rumbold and Pierscionek (2018).

Whether stop lists or allow lists are preferable depends on the context of application. Allow lists are more restrictive and require less effort from the analysts, by automatically excluding all terms, times, locations, etc. that are not explicitly considered beforehand. For the spatial context, for instance, unless worldwide data is required, allow lists are frequently used, to limit data collection to a specific area, region, place, etc. Conversely, stop lists can be added selectively on top, to exclude places that are known to be related to vulnerable groups or sensitive contexts (e.g., hospitals, party locations). Similarly, filter lists for specific terms, hashtags, or emoji can be defined for the topical context.

For topical contexts, the openness of possible references complicates defining holistic stop lists ahead of time. As an example, Fig. 14.4 shows a map generated from terms, hashtags, and emoji used on the social media services Twitter, Flickr, and Instagram at a public vantage point and park. The syringe emoji could indicate drug use, which may lead to further onsite investigation by, e.g., authorities, with potential unexpected consequences of the user perspective. Obviously, this is an edge case for social-individual privacy because both positive (society) and negative (user) consequences are imaginable. One solution would be to assign the specific emoji to a thematic broader *emoji class*, e.g., the umbrella group of "medical emoji"[1] (see Sect. 14.3.2). As another solution, the syringe emoji could be classified ahead of time, for increased sensitivity, leading to, e.g., a greater spatial granularity reduction on data ingestion, or exclusion, preventing having to deal with this ambiguous ethical edge case in advance.

Lastly, as the second approach to enable systematic user privacy with HLL, threshold values may be defined, similar to what is known from other disciplines, such as the HIPAA Privacy Rules for health data publications (Malin et al. 2011) or census statistics (Szibalski 2007, p.142). Allshouse et al. (2010), for instance, use geomasking in combination with k-anonymity, to define a lower threshold of $k = 5$ (people), which is a rule of thumb size in geoprivacy (Kamp et al. 2013). Comparable best-practice threshold values could be defined for HLL sets of different sizes, e.g., suggestions by Desfontaines et al. (2019), with smaller sets indicating lesser privacy protection due to a scarce context collapse. In the spatial context, this could be implemented by using quadtrees, for example, to split and aggregated social data into sub-sections (quads), based on pre-defined thresholds, where the resolution is automatically decreased for areas of lesser data density.

[1] Unicode Consortium, unicode.org/emoji/charts-13.0/full-emoji-list.html#medical.

Table 14.2 Example of sensitive context factors for which no data analysis might be carried out

Type of context	Example of sensitive context factor	Reference
Spatial context	– Home location	(Georgiadou et al. 2019), (Kim et al. 2021)
	– Hospitals	(Ağır et al. 2016), (Kim and Kwan 2021)
	– Related to specific events (concert grounds, party locations)	(Such et al. 2017)
Temporal context	– Nighttimes	(Nikas et al. 2018)
	– Past and archived content, time collapse	(Brandtzaeg and Lüders 2018)
	– During specific events (e.g., new year, Thanksgiving, 4th of July)	(Such et al. 2017)
Topical context	– Activists, protesters, dissidents	(Uldam 2018)
	– Health issues (e.g., related to diabetes or corona)	(Matković et al. 2021)
Social context	– Children	(Steinberg 2016), (Marwick and Boyd 2014)
	– LGBTQ+[a]	(Birnholtz et al. 2020)
	– Personal, social relationships	(Houghton and Joinson 2010)
	– Minorities (race and religion)	(Mashhadi et al. 2021)

[a] Lesbian, gay, bisexual, transgender, queer, and others

Fig. 14.4 A thematically sensitive emoji on drug use at selected locations

14.4 Case Study

Even though different implementations of HLL exist, all share a number of basic steps. At the core, the binary representation of any given character string is divided into *buckets*, for which the number of leading zeroes is counted (see Sect. 14.2.4). Because any given character string is first randomized, it is possible to predict how many distinct items must have been added to a given HLL set, based on the maximum number of leading zeroes observed. In other words, if multiple items are added to an HLL set, only the highest number of leading zeroes per bucket needs to be memorized. As a result, the cardinality estimation will only approximate counts.

As a side effect, there is a limited ability to check whether a specific user or ID has been added to a HLL set. In an adversarial situation, Desfontaines et al. (2019) refer to such a check as an *intersection attack*. Intersection attacks first require obtaining the hash of a targeted person or ID and then adding this hash to an HLL set. If the HLL set changes, an adversarial may be able to increase their initial suspicion by a certain degree. To better illustrate intersection attacks and how and under which circumstances the privacy of a user could become compromised in the presented two-component research setup, we briefly introduce two examples.

Alex is included in the YFCC100M dataset (Thomee et al. 2016) because he published 289 photos under Creative Commons Licenses between 2013 and 2014 on Flickr; 120 of these photos are geotagged. Given this information, it will be relatively easy to re-identify Alex. Sandy is an internal adversary. She could be someone working at an analytics service with full access to the database. Robert, on the other hand, is someone representing an external adversary, with access only to the published dataset. In the first example, the privacy of Alex is compromised if Sandy could increase or confirm her suspicion that Alex was not at his workplace in Berlin on 9 May 2012. In the second example, the privacy of Alex is compromised if Robert could increase or confirm his suspicion that Alex was indeed at least once at a specific location, e.g., contrary to what Alex claims. Finally, Alex could be

someone who voluntarily contributed his pictures to the conceived analytics service or altruistically published Creative Commons photos on Flickr.

Consider that, at the moment of contribution, Alex may not have thought of the consequences for his privacy but later realized his mistake. With the use of raw data, even removing any compromising data from Flickr, this change would need to be reflected in any subsequent data collection, such as in the analytics service or the YFCC100M dataset. This is either impractical or impossible. The question is, therefore, whether it is possible to replace raw data workflows with a privacy-aware visualization pipeline, without significantly reducing utility.

Several factors must coincide for intersection attacks to be successful. Firstly, an adversarial must have access to HLL sets. In our system model, this can either be an internal adversary (Sandy), having direct access to the database, or an external adversary (Robert), having access only to published data. Furthermore, an adversary must be able to either compute hashes for a given target user or somehow gain access to a computed HLL set for the given user. The former is only possible if the secret key is compromised. The latter appears conceivable, in our example, if the adversary has some prior knowledge about other locations visited by a target user, and if the HLL sets of these locations ideally contain only the target user or a few other users. In the following, we explore this worst-case scenario, where both Sandy and Robert somehow got hold of an HLL set that only contains Alex's computed hashes.

For Sandy, this means in order to test whether Alex was not in Berlin on 9 May 2012, she either needs Alex's original user ID and the secret key to construct the hash or find another location that has only been visited by Alex on this date. In this unlikely scenario, the result of an intersection attack for all grid cells is shown in Fig. 14.5. Visible in the figure is that a large number of other grid cells show false positives for the intersection test, that is, these HLL sets did not change, even when updated with the particular user day-hash for Alex.

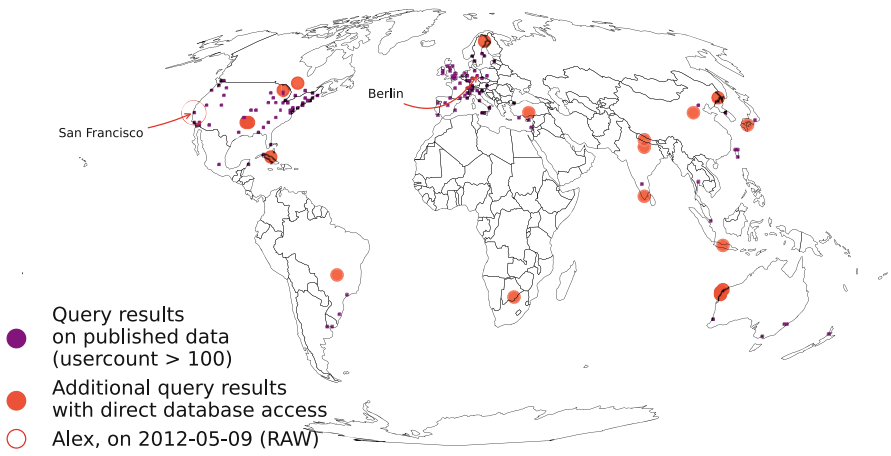

Fig. 14.5 Evaluation of scenario "Sandy" (Dunkel et al. 2020, CC-BY 4.0)

Since HLL prevents the occurrence of false negatives, and San Francisco is indeed among these locations, the result does include Alex's actual location on 9 May 2012. Depending on the size of the targeted HLL set, Sandy may then increase her suspicion by some degree. In the case of the grid cell for San Francisco, with 209,581 user days, this increase in posterior knowledge may be found to be negligibly small. In other words, even if there was no post from Alex on 9 May 2012, the intersection attack may have produced the same result. In conclusion, even in the worst scenario, having direct access to the database and a compromised secret key, Sandy could not gain any further affirmation.

Similarly, and rather incidentally, the positive grid cell for Berlin does indeed falsely suggest that Alex was in Berlin. This is not surprising given that larger HLL sets have a higher likeliness of showing false positives and Berlin is a highly frequented location. In other words, Alex benefits from the privacy-preserving effect of HLL.

In the second scenario, consider a situation in which Robert may have an a priori suspicion that Alex went to Cabo Verde. Alex, on the other hand, does not want Robert to know that he went surfing without him. Robert knows that Alex is participating in the conceived analytics service and, somehow, gains access to an HLL set containing only one hashed user ID from Alex. The results of the intersection attack for all grid cells are shown in Fig. 14.6. Since only 56 users have been to Cabo Verde in the YFCC100M dataset, the particular bin is not included in the published benchmark data, which is limited by a minimum threshold of 100 users. However, with direct access to the database, Robert could observe that Cabo Verde is among the locations revealed. In this case, Robert may gain some affirmation for his suspicion that Alex was in Cabo Verde. At the same time, a definite answer will not be possible, given the irreversible approximation of the HLL structure. For example, for the same intersection attack, for set sizes below 56

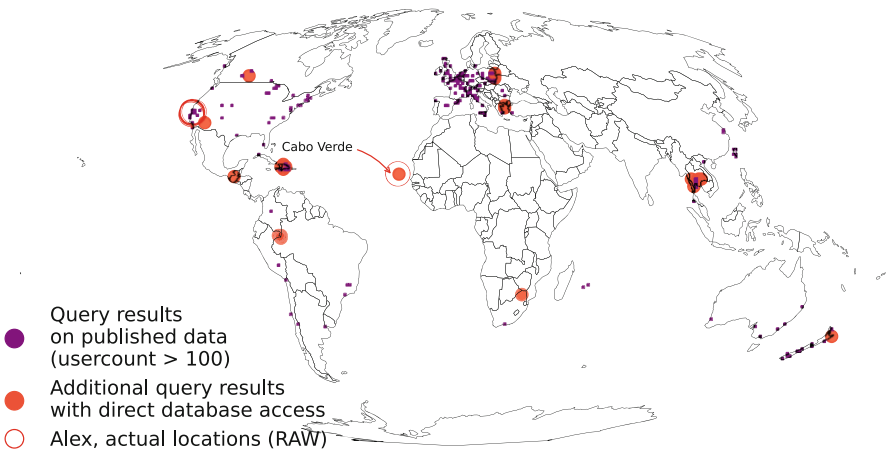

Fig. 14.6 Evaluation of scenario "Robert" (Dunkel et al. 2020, CC-BY 4.0)

users, there are 14 other grid cells that show false positives, down to 8 users. In other words, even though these HLL sets do not change when tested, Alex has never been to these locations.

While these two scenarios provide a base to understand how intersection attacks may be executed in a spatial setting, a valid question is how likely successful intersection attacks are overall. To some degree, this depends on questions of security, such as protecting the secret key or managing database access.

Another part is directly related to the distribution of collected data and the number of outliers that are present at each stage of data processing. If data is more clustered, users will generally receive more benefits from the privacy-preserving effects of HLL. This can be quantitatively substantiated with the given dataset (Dunkel et al. 2020).

14.5 Conclusion

The research presented in this chapter introduced a number of approaches to deal with privacy aspects in the process of social media data processing. Social media data is being used as a source of data for wide-ranging projects within and beyond the scope of this book (see Sect. 14.2.1). The relevance of privacy aspects in processing this kind of data and the range of related work are pointed out in Sect. 14.2.2. Furthermore, in Sect. 14.2.3, we discussed our focus on data retention as a potential threat for analysts. We defined the term and explained that we make use of that specific term to emphasize the explosiveness of dealing with personal data.

We showed that it is possible to preserve the privacy of social media users with the major concepts. As a basis for our first concept, we first introduced the cardinality estimation algorithm HyperLogLog in Sect. 14.2.4. In Sect. 14.3.1, the main part of this chapter, we introduced a concept to store social media data in a way that it is not possible to extract individual items from it but only to estimate the cardinality of social media data items within a certain set, plus running set operations over multiple sets to extend analytical ranges. Applying this method requires defining the scope of the result before even gathering the data and thus prevents the data from being misused for other purposes at a later point in time. This follows the *privacy-by-design* principle.

As an extension to the first concept, we proceeded by introducing a concept that is well known in the geographic community, generalization, in Sect. 14.3.2. By defining a number of abstraction layers, it is possible to even more reduce the data to be stored, depending on the required precision. The less precise data is needed, the fewer data needs to be stored. Finally, in Sect. 14.3.3, we explain the conceptual exclusion of edge cases by applying filter lists to the data set.

A closing case study in Sect. 14.4 explains the concept of intersection attacks and shows that under rare circumstances the HyperLogLog technology is vulnerable against them. The case study unveils that the larger the dataset, the less likely are

intersection attacks. Since social media data is usually very large, implementing the HyperLogLog technology is an excellent approach to protect the data from being abused, thieving, or publicly exposed and thus preserves the privacy of social media users.

Acknowledgments This research was supported by the German Research Foundation DFG within Priority Research Program 1894 *Volunteered Geographic Information: Interpretation, Visualization and Social Computing* (VGIscience, EVA-VGI, BU 2605/8-2).

References

Ağır B, Huguenin K, Hengartner U, Hubaux J-P (2016) On the privacy implications of location semantics. In: Proceedings on Privacy Enhancing Technologies. https://doi.org/10.1515/popets-2016-0034

Allshouse WB, Fitch MK, Hampton KH, Gesink DC, Doherty IA, Leone PA, Serre ML, Miller WC (2010) Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. Geocarto Int, 443–452. https://doi.org/10.1080/10106049.2010.496496

Almås I, Attanasio O, Jalan J, Oteiza F, Vigneri M (2018) Using data differently and using different data. J Dev Eff, 462–481. https://doi.org/10.1080/19439342.2018.1530279

Büchi M, Festic N, Latzer M (2022) The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. Big Data Soc. https://doi.org/10.1177/20539517211065368

Beresford AR, Stajano F (2003) Location privacy in pervasive computing. IEEE Pervasive Comput, 46–55. https://doi.org/10.1109/mprv.2003.1186725

Berghel H (2018) Malice domestic: The cambridge analytica dystopia. Computer, 84–89. https://doi.org/10.1109/mc.2018.2381135

Birnholtz J, Kraus A, Zheng W, Moskowitz DA, Macapagal K, Gergle D (2020) Sensitive sharing on social media: Exploring willingness to disclose prep usage among adolescent males who have sex with males. Soc Media Soc. https://doi.org/10.1177/2056305120955176

Blanchette J-F, Johnson DG (2002) Data retention and the panoptic society: The social benefits of forgetfulness. Inf Soc, 33–45. https://doi.org/10.1080/01972240252818216

Boyd D, Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc, 662–679. https://doi.org/10.1080/1369118x.2012.678878

Brandtzaeg PB, Lüders M (2018) Time collapse in social media: extending the context collapse. Soc Media Soc. https://doi.org/10.1177/2056305118763349

Burghardt D, Duchêne C, Mackaness W (2016) Abstracting geographic information in a data rich world. Springer, New York. https://doi.org/10.1007/978-3-319-00203-3

Cavoukian A et al. (2009) Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario, Canada

Ciriani V, Di Vimercati SDC, Foresti S, Samarati P (2007) κ-anonymity. In: Secure data management in decentralized systems. Springer, New York, pp 323–353. https://doi.org/10.1007/978-0-387-27696-0_10

Cosenza V (2022) World map of social networks. https://vincos.it/world-map-of-social-networks. Accessed 19-Jul-2022

Daly, A, Devitt, SK, Mann, M (2019) Good Data, Theory on Demand, 29. Institute of Network Cultures, Amsterdam. https://networkcultures.org/blog/publication/tod-29-good-data/

Desfontaines D, Pejó B (2020) Sok: differential privacies. In: Proceedings on Privacy Enhancing Technologies, pp 288–313. https://doi.org/10.2478/popets-2020-0028

Desfontaines D, Lochbihler A, Basin D (2019) Cardinality estimators do not preserve privacy. In: Proceedings on Privacy Enhancing Technologies, pp 26–46. https://doi.org/10.2478/popets-2019-0018

Dowling M-E (2022) Cyber information operations: Cambridge analytica's challenge to democratic legitimacy. J Cyber Policy, 1–19. https://doi.org/10.1080/23738871.2022.2081089

Duckham M, Kulik L (2005) A formal model of obfuscation and negotiation for location privacy. In: International Conference on Pervasive Computing. Springer, pp 152–170. https://doi.org/10.1007/11428572_10

Dunkel A (2021) Tag maps in der Landschaftsplanung. Springer Fachmedien Wiesbaden, Wiesbaden, pp 137–166. https://doi.org/10.1007/978-3-658-29862-3_8

Dunkel A, Andrienko G, Andrienko N, Burghardt D, Hauthal E, Purves R (2019) A conceptual framework for studying collective reactions to events in location-based social media. Int J Geogr Inf Sci, 780–804. https://doi.org/10.1080/13658816.2018.1546390

Dunkel A, Löchner M, Burghardt D (2020) Privacy-aware visualization of volunteered geo-graphic information (vgi) to analyze spatial activity: A benchmark implementation. ISPRS Int J Geo-Inf. https://doi.org/10.3390/ijgi9100607

Dwork C (2008) Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, pp 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

European Commission (2018) What is personal data? https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data. Accessed 21-Nov-2022

Fiedrich F, Fathi R (2021) Humanitäre hilfe und konzepte der digitalen hilfeleistung. In: Sicherheitskritische Mensch-Computer-Interaktion. Springer, pp 539–558. https://doi.org/10.1007/978-3-658-32795-8_25

Flajolet P, Fusy E, Gandouet O, Meunier F (2007) Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. Discrete Math Theor Comput Sci. https://doi.org/10.46298/dmtcs.3545. https://dmtcs.episciences.org/3545

Georgiadou Y, de By RA, Kounadi O (2019) Location privacy in the wake of the gdpr. ISPRS Int J Geo-Inf. ISSN 2220-9964. https://doi.org/10.3390/ijgi8030157

Guillou C, Portner C (2020) Data retention - more than meets the eye. https://www.theprivacyhacker.com/2020/12/data-retention/

Hauser C, Kabatnik M (2001) Towards privacy support in a global location service. In: Proceedings of the IFIP Workshop on IP and ATM Traffic Management, pp 81–89

Hearst MA, Pedersen E, Patil L, Lee E, Laskowski P, Franconeri S (2019) An evaluation of semantically grouped word cloud designs. IEEE Trans Vis Comput Graph, 2748–2761. https://doi.org/10.31219/osf.io/3eutf

Hildebrandt M (2006) Privacy and identity. Privacy and the criminal law. Intersentia, Antwerp/Oxford

Houghton DJ, Joinson AN (2010) Privacy, social network sites, and social relations. J Technol Hum Serv, 74–94. https://doi.org/10.1080/15228831003770775

Jiang B (2013) Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. Prof Geogr, 482–494. https://doi.org/10.1080/00330124.2012.700499

Kamp M, Kopp C, Mock M, Boley M, May M (2013) Privacy-preserving mobility monitoring using sketches of stationary sensor readings. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp 370–386. https://doi.org/10.1007/978-3-642-40994-3_24

Keßler C, McKenzie G (2018) A geoprivacy manifesto. Trans GIS. https://doi.org/10.1111/tgis.12305

Kherwa P, Bansal P (2019) Topic modeling: a comprehensive review. EAI Endors Trans Scal Inf Syst. https://doi.org/10.4108/eai.13-7-2018.159623

Kim J, Kwan M-P (2021) An examination of people's privacy concerns, perceptions of social benefits, and acceptance of covid-19 mitigation measures that harness location information: A

comparative study of the us and south korea. ISPRS Int J Geo-Inf, 25. https://doi.org/10.3390/ijgi10010025
Kim J, Kwan M-P, Levenstein MC, Richardson DB (2021) How do people perceive the disclosure risk of maps? Examining the perceived disclosure risk of maps and its implications for geoprivacy protection. Cartogr Geogr Inf Sci, 2–20. https://doi.org/10.1080/15230406.2020.1794976
Kounadi O, Resch B, Petutschnig A (2018) Privacy threats and protection recommendations for the use of geosocial network data in research. Soc Sci, 191. https://doi.org/10.3390/socsci7100191
Kuner C, Marelli M (2020) Data analytics and big data. International Committee of the Red Cross, Geneva, Switzerland, pp 92–111
Löchner M, Fathi R, Schmid D, Dunkel A, Burghardt D, Fiedrich F, Koch S (2020) Case study on privacy-aware social media data processing in disaster management. ISPRS Int J Geo-Inf, 709. ISSN 2220-9964. https://doi.org/10.3390/ijgi9120709
Leetaru K (2019) The era of precision mapping of social media is coming to an end. https://www.forbes.com/sites/kalevleetaru/2019/03/06/the-era-of-precision-mapping-of-social-media-is-coming-to-an-end/
Maireder A, Schlögl S, Schütz F, Karwautz M, Waldheim C (2014) The european political twittersphere: Network of top users discussing the 2014 european elections. University of Vienna, Viena
Malin B, Benitez K, Masys D (2011) Never too old for anonymity: a statistical standard for demographic data sharing via the hipaa privacy rule. J Am Med Inf Assoc, 3–10. https://doi.org/10.1136/jamia.2010.004622
Manokha I (2018) Surveillance, panopticism, and self-discipline in the digital age. Surveillance Soc, 219–237. https://doi.org/10.24908/ss.v16i2.8346
Marwick AE, Boyd D (2014) Networked privacy: How teenagers negotiate context in social media. New Media Soc, 1051–1067. https://doi.org/10.1177/1461444814543995
Mashhadi A, Winder SG, Lia EH, Wood SA (2021) No walk in the park: The viability and fairness of social media analysis for parks and recreational policy making. In: ICWSM, pp 409–420. https://doi.org/10.1609/icwsm.v15i1.18071
Matković R, Vejmelka L, Ključević Ž (2021) Impact of covid 19 on the use of social networks security settings of elementary and high school students in the split-dalmatia county. In: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, pp 1476–1482. https://doi.org/10.23919/mipro52101.2021.9597179
Miller V (2020) Understanding digital culture. SAGE Publications Limited, London, UK
Mooney P, Olteanu-Raimond A-M, Touya G, Juul N, Alvanides S, Kerle N (2017) Considerations of privacy, ethics and legal issues in volunteered geographic information. Map Citizen Sensor, 119–135. https://doi.org/10.5334/bbf.f
Morton GM (1966) A computer oriented geodetic data base and a new technique in file sequencing. International Business Machines Company, New York
Niemeyer G (2008) geohash.org is public! https://blog.labix.org/2008/02/26/geohashorg-is-public. Accessed 06-Sep-2022
Nikas A, Alepis E, Patsakis C (2018) I know what you streamed last night: On the security and privacy of streaming. Digit Investig, 78–89. https://doi.org/10.1016/j.diin.2018.03.004
Reyman J (2013) User data on the social web: Authorship, agency, and appropriation. Coll Engl, 513–533
Rucz M, Kloosterboer S (2020) Data retention revisited. https://edri.org/our-work/launch-of-data-retention-revisited-booklet/
Rumbold JM, Pierscionek BK (2018) What are data? A categorization of the data sensitivity spectrum. Big Data Res, 49–59. https://doi.org/10.1016/j.bdr.2017.11.001
Solove DJ (2008) Understanding privacy. Harvard University Press, Cambridge, MA
Steinberg SB (2016) Sharenting: Children's privacy in the age of social media. Emory LJ, 839
Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics–challenges in topic discovery, data collection, and data preparation. Int J Inf Manag, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Such JM, Porter J, Preibusch S, Joinson A (2017) Photo privacy conflicts in social media: A large-scale empirical study. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17. Association for Computing Machinery, New York, NY, USA, pp 3821–3832. ISBN 9781450346559. https://doi.org/10.1145/3025453.3025668

Szibalski M (2007) Textteil - Kleinräumige Bevölkerungs- und Wirtschaftsdaten in der amtlichen Statistik Europas. Wirtschaft und Statistik, 137–143

Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li L-J (2016) Yfcc100m: The new data in multimedia research. Commun ACM, 64–73. https://doi.org/10.1145/2812802

Uldam J (2018) Social media visibility: challenges to activism. Media Cult Soc, 41–58. https://doi.org/10.1177/0163443717704997

Wang Z, Ye X (2018) Social media analytics for natural disaster management. Int J Geogr Inf Sci, 49–72. https://doi.org/10.1080/13658816.2017.1367003

Williams ML, Burnap P, Sloan L (2017) Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. Sociology, 1149–1168. https://doi.org/10.1177/0038038517708140

Wilson S (2020) The era of antisocial social media. https://hbr.org/2020/02/the-era-of-antisocial-social-media

Wright, DN, Demetres, MR, Mages, KC, DeRosa, AP, Jedlicka C, Stribling JC, Baltich Nelson B, Delgado, D (2020) How long should we keep data? An evidence-based recommendation for data retention using institutional meta-analyses. Samuel J. Wood Medical Library: Faculty Publications

Zhou B, Pei J, Luk W (2008) A brief survey on anonymization techniques for privacy preserving publishing of social network data. In: ACM Sigkdd Explorations Newsletter, pp 12–22. https://doi.org/10.1145/1540276.1540279