Protecting privacy using HyperLogLog to process data from Location Based Social Networks

marc.loechner@tu-dresden.de
alexander.dunkel@tu-dresden.de
dirk.burghardt@tu-dresden.de

A large share of the content in Social Media services is publicly available. Its location data combined with other information can be utilized e.g. to draw maps based on trending topics or social groups. Users may not be aware of this third-party data processing and regard their privacy violated.

In this paper we propose a concept to protect privacy for users of Location Based Social Network services, when processing their shared data. We achieve this by applying an algorithm called HyperLogLog (HLL) to the processing of LBSN data. The concept is designed on behalf of good intentions and to prevent *accidental disclosure* by actors who are aware of privacy issues.

Introduction

Location Based Social Networks (LBSN), such as Twitter or Instagram, provide a rich set of data. The primary motivation of users on these networks is to communicate and share information. However, since this data is publicly available on the internet, it can be used for applications beyond the LBSN service itself. With increasing network speeds and device performance, processing of LBSN data will soon be possible even in real time. It is about time to think of and prepare for the consequences.

The main problem with LBSN data utilization for applications other than their dedicated use case is that explicit consent from the LBSN 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, may that be scientific, commercial or administrative (Boyd & Crawford, 2012). In the specific case of LBSN data, there is a particular focus on aspects related to privacy. In contrast to other environments, the data to be protected is already public (Williams, Burnap, & Sloan, 2017).

In the view of a malicious actor, that data is already “compromised” and therefore can be utilized for any purpose, including those that oppose the user’s interest (Zhou, Pei, & Luk, 2008). But data can also be used with good intentions (Daly, Devitt, & Mann, 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 (Bosch et al., 2011).

An increasing skepticism and fear that data is misused may motivate LBSN users stop using LBSN services and retreat to closed groups like Instant Messaging (IM) services (Leetaru, 2019). In order to support ongoing development of positive use cases, scientists need to respect and actively protect LBSN 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 to the use of LBSN data following the *privacy by design* principles (Cavoukian & others, 2009), by making use of multiple layers of abstraction.

In this paper we show a concept that achieves privacy by implementing an algorithm called HyperLogLog (HLL) (Flajolet, Fusy, Gandouet, & Meunier, 2007) and applying it to LBSN data. The key aspect for our *privacy by design* approach here is to make it impossible to relate to the original LBSN data from a given processed data set.

The objectives for this research are to apply HLL to the process of analysis and visualization of LBSN data, on behalf of good intentions, and to prevent *accidental disclosure* by actors who are aware of the privacy aspects.

Related work

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 (Cavoukian & others, 2009) is a relevant objective in the conception of applications in general. Concepts that are built upon these principles are hard to break in terms of privacy violations.

A wide number of approaches have dealt with technical methods to protect privacy. A general method to anonymize data has been presented as *differential privacy* (DP) (Dwork, 2008) and adopted frequently (Desfontaines & Pejó, 2019). In its core, differential privacy adds noise to data sets to protect the individual data when being queried. 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 and thus, makes it hard to apply on LBSN data.

In the geo-community, there are a number of concepts to protect the privacy in

terms of location data. Several techniques have been introduced that are based on anonymity (*mix zones* (Beresford & Stajano, 2003), *k-anonymity* (Ciriani, Di Vimercati, Foresti, & Samarati, 2007)), obfuscation (*imprecision* (Duckham & Kulik, 2005)) or policy (*restriction* (Hauser & Kabatnik, 2001)). All of these also require the possession of original raw data. The processed data sets are unable to be updated with subsequent data, which requires reprocessing of the entire data set upon updates. This is again very inefficient when dealing with LBSN data.

In the context of VGI, the consideration of privacy, ethics and legal issues should play an important role, however, so far only a few researchers have dealt with it. The statement from Mooney et al. (2017): “Privacy of user data and information should be considered in the initial design of VGI systems” can be extended to platforms and methods for the analysis and further processing of VGI and LBSN.

Kounadi et al. (2018) discusses privacy risks related to the analysis of *geosocial media data* and provides *geoprivacy-by-design* recommendations for sharing this data and publishing resulting maps. e.g. reduce the spatial and/or temporal resolution of public maps or consider the use of heat maps.

21 theses are formulated by Keßler and McKenzie (2018) to reflect on the current state of *geoprivacy* from a technological, ethical, legal and educational perspective. They provide various examples how common it has become to share location and how it can be used and misused.

The concept of *abstraction* has been widely used in the geo-community to visualize spatial information scale dependent with different degrees of detail (Burghardt, Duchêne, & Mackaness, 2016). We re-dedicate these generalization methods from geo-visualization to privacy protection. Therefore we have introduced a conceptual model to protect privacy of LBSN users. We have eliminated precise data by deriving multiple abstraction layers from it to be able to quantitatively describe different levels of privacy (Löchner, Dunkel, & Burghardt, 2018).

Concept

Processing LBSN data often means counting events and interpreting the result. For example, a query could get all posts from within a certain area during a certain time period that include a certain hashtag. The result will be a list of posts that match the requested criteria.

This set of data can then be used to e.g. draw a map that highlights areas with postings about the topic, soon likely even in real time. The problem regarding privacy about this set of data is, that it is easy to use the data for other purposes than the intended, e.g. get all the user names of these posts, and e.g. discover what else they post about. This is possible because once you have a full set of data you can do more with it than just counting its elements.

To approach that problem, we propose a method to process LBSN data that utilizes an algorithm called HyperLogLog (HLL) (Flajolet et al., 2007). HLL does not store the original data, but a structure of hashes. This structure is called a *shard*.

First every element of a query is hashed. Then an estimation over the entire set decides whether an item is already stored or new. In case of the latter, the hash is added to the shard. It is now impossible to retrieve the original values of the elements.

Referring to the list of posts mentioned above: every post in the queried list will be hashed and stored in the shard. Now it is impossible to read every post in the shard, but only to determine the number of them. Many different queries can be stored in parallel, e.g. for different locations, topics, times, etc. (Fig. 1).

Generally speaking, this means that every shard can only answer one question, e.g. how many posts are within a certain area during a certain time period that include a certain hashtag. This makes the shard “disposable” in a sense. On the other hand, it complies with the principle of *privacy by design*, because it is impossible to gain other knowledge from the data than which it was intended to.

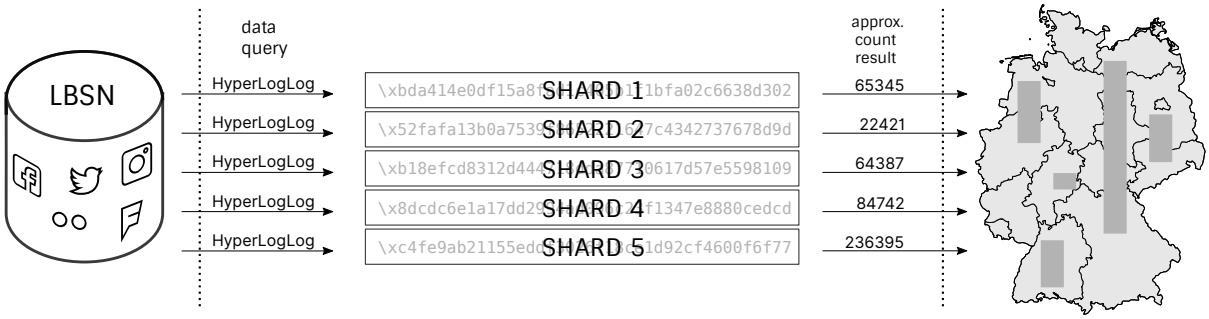

Fig. 1: HyperLogLog data processing flow: query, hashes, count result, visualization

In contrast to just storing the number of elements as a plain integer, using HLL it is possible to check whether a certain element is already in the shard, and only add it, if not. This is crucial for real time processing of LBSN data, because it makes adding upcoming posts that match the requested criteria to the shard possible.

Furthermore, using HLL on LBSN data allows a number of set theory operations on shards, for example combining multiple shards and calculate cardinality over a *union* of them. This allows for example to count the total number of common visitors over two separate locations. A union of multiple shards can create higher abstraction layers (Löchner et al., 2018), e.g. combine postings of places in a city, and create a *mix zone* which improves privacy protection (Beresford & Stajano, 2003).

Similarly, it is possible to *intersect* multiple HLL shards to count common patterns. This can be useful to identify a number of visitors of both locations. However, intersection is not very reliable if the shards have very few overlap or a large difference in size. This again is beneficial to privacy, since this limitation prevents identifying outliers or *sensitive cells* (Zhou et al., 2008) from an attacker’s viewpoint.

Data storage and processing speed is a major issue when dealing with big data (Cano, 2014). Using HLL increases the process significantly compared to conventional analytics of the original data (Othman, 2018). The storage volume of

HLL data breaks down to only 0.5% of the original data, while maintaining about 2% approximate error rate (Flajolet et al., 2007).

Discussion

The goal of our research is to propose methods to process LBSN data that follow *privacy by design* principles (Cavoukian & others, 2009). The contribution focuses on providing a reduced data set that does not include the original raw data, but an estimated distinct count of each set of information.

An advantage of HLL for the protection of privacy is the ability to compute the HLL shards directly within a query to the original data, before storing anything to a local database. This way, an operator never gets in possession of the original data. The ability of HLL shards to be computed with methods from the set theory like *union* and *intersection* makes them valuable for more advanced analytics and is an advantage over storing just plain numbers.

Further advantages of HLL are the very high processing speed of large amounts of data and the very low storage space, compared to working with raw data (Flajolet et al., 2007). The latter may be regarded as tempered by the drawback of the data being only useful for one task.

The primary drawbacks of HLL data is that more precise planning of the design of the data structure is required. One needs to know in advance, what exactly is to be counted. For every information, a new query to the original data is necessary, and for every new data input all involved HLL shards will be updated. Aside from that, it is also not possible to delete single entries from a shard. Since HLL data is considered as statistical data, this is an expected drawback and does not pose a barrier to privacy protection.

Desfontaines et al. (2019) state, that algorithms like HLL do not preserve privacy. This of course depends on the attacker model. Their example assumes the original data to be secretly stored and the attacker gains knowledge of some of it by guessing. Our concept in contrast targets the processing of data, that is already publicly available on the internet. It addresses data operators, who want to proactively prevent the *accidental disclosure* of LBSN users e.g. in visualization, by storing their data in a data structure, where only exactly that data can be read from, that is required for the task.

In future work, we will provide more detailed technical explanation of the methodology and show an implementation of this concept. We will provide a reference implementation based on standard software and documentation. Furthermore, we are planning to provide sample case studies based on the YFCC100M data set (2015).

Conclusion

In this paper we have proposed a methodological concept to accomplish privacy

protection of LBSN users, when processing LBSN data. With a focus on the *privacy by design* principle, our approach was to make access to the original LBSN data from a given processed data set impossible. We achieved this by applying an algorithm called HyperLogLog (HLL) into the process of analysis and visualization of LBSN data. The resulting framework acts as a proof-of-concept for the proposed method to protect privacy of LBSN data. It is designed on behalf of good intentions and to prevent *accidental disclosure* by actors who are aware of the privacy aspects.

References

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive Computing*, (1), 46–55.

Bosch, H., Thom, D., Wörner, M., Koch, S., Püttmann, E., Jäckle, D., & Ertl, T. (2011). Scatterblogs: Geo-spatial document analysis. *2011 ieee conference on visual analytics science and technology (vast)*, 309–310. IEEE.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Burghardt, D., Duchêne, C., & Mackaness, W. (2016). *Abstracting geographic information in a data rich world*. Springer.

Cano, J. (2014). *The v's of big data: Velocity, volume, value, variety, and veracity*. Retrieved from https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity

Cavoukian, A., & others. (2009). Privacy by design: The 7 foundational principles. *Information and Privacy Commissioner of Ontario, Canada*, *5*.

Ciriani, V., Di Vimercati, S. D. C., Foresti, S., & Samarati, P. (2007). κ -anonymity. In *Secure data management in decentralized systems* (pp. 323–353). Springer.

Daly, A., Devitt, S. K., & Mann, M. (2019). *Good data*. Institute of Network Cultures.

Desfontaines, D., Lochbihler, A., & Basin, D. (2019). Cardinality estimators do not preserve privacy. *Proceedings on Privacy Enhancing Technologies*, *2019*(2), 26–46.

Desfontaines, D., & Pejó, B. (2019). SoK: Differential privacies. *arXiv Preprint arXiv:1906.01337*.

Duckham, M., & Kulik, L. (2005). A formal model of obfuscation and negotiation for location privacy. *International conference on pervasive computing*, 152–170. Springer.

Dwork, C. (2008). Differential privacy: A survey of results. *International conference on theory and applications of models of computation*, 1–19. Springer.

Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007). Hyperloglog: The analysis of a

near-optimal cardinality estimation algorithm. *Discrete mathematics and theoretical computer science*, 137–156. Discrete Mathematics; Theoretical Computer Science.

Hauser, C., & Kabatnik, M. (2001). Towards privacy support in a global location service. *Proceedings of the ifip workshop on ip and atm traffic management*, 81–89.

Hildebrandt, M. (2006). Privacy and identity. *Privacy and the Criminal Law*, *43*.

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, *22*.

Kounadi, O., Resch, B., & Petutschnig, A. (2018). Privacy threats and protection recommendations for the use of geosocial network data in research. *Social Sciences*, *7*(10), 191.

Leetaru, K. (2019). The era of precision mapping of social media is coming to an end. Retrieved from Forbes website: https://www.forbes.com/sites/kalevleetaru/2019/03/06/the-era-of-precision-mapping-of-social-media-is-coming-to-an-end

Löchner, M., Dunkel, A., & Burghardt, D. (2018). A privacy-aware model to process data from location-based social media. *VGI geovisual analytics workshop, colocated with bdva 2018*.

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. *Mapping and the Citizen Sensor*, 119–135.

Othman, A. (2018). *Social media data mining and analytics* (pp. 217–223).

Reyman, J. (2013). User data on the social web: Authorship, agency, and appropriation. *College English*, *75*(5), 513–533.

Solove, D. J. (2008). *Understanding privacy* (Vol. 173). Harvard university press Cambridge, MA.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., … Li, L.-J. (2015). YFCC100M: The new data in multimedia research. *arXiv Preprint arXiv:1503.01817*.

Williams, M. L., 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*, *51*(6), 1149–1168.

Zhou, B., Pei, J., & Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, *10*(2), 12–22.