Big Data Privacy and Anonymization

Vicenç Torra, Guillermo Navarro-Arribas

Big data privacy and anonymization

Vicenç Torra[1] Guillermo Navarro-Arribas[2]

[1] School of Informatics,
University of Skövde, Sweden
[2] Department of Information and Communication Engineering,
Universitat Autònoma de Barcelona, Catalonia, Spain
Email: vtorra@his.se, guillermo.navarro@uab.cat

Abstract. Data privacy has been studied in the area of statistics (statistical disclosure control) and computer science (privacy preserving data mining and privacy enhancing technologies) for at least 40 years. In this period models, measures, methods, and technologies have been developed to effectively protect the disclosure of sensitive information.
The coming of big data, with large volumes of data, dynamic and streaming data, poses new challenges to the field. In this paper we will review some of these challenges and propose some lines of research in the field.

1 Introduction

Data privacy studies models and methods to ensure that there is no disclosure of sensitive information. The field arose within the statistics community to ensure that sensitive data from census were not disclosed. Later, the problem appeared within the computer science community to ensure privacy in communications, and databases. Three main research communities exist today: statistical disclosure control, privacy enhancing technologies, and privacy preserving data mining. They study similar problems, although the focus is slightly different due to the types of data they consider and the type of uses of these data.

The field has now more than 40 years, starting with e.g. the seminal papers of Dalenius [5, 6], Chaum [4], and Denning and Schlöder [7]. During these years, different types of privacy models have been defined, methods to protect sensitive information according to these privacy models have been proposed, and measures for evaluating disclosure risk, and information loss have also been defined. There is a large number of approaches for different types of data. This does not mean that all problems are solved, but there exists already a solid and useful set of techniques for ensuring different levels of privacy for some types of applications. See e.g. the reference books [8, 11, 19] for details.

The increasing amount of information available, and the coming of big data and data science poses new problems to the field. In this paper we will review some of these problems, and outline accordingly some lines for further research.

The new EU General Data Protection Regulation includes the implementation of the right to rectification and the right to be forgotten. That is, companies need to modify or delete records from a database when users and citizens want to take advantage of these rights. In order to implement these rights, data provenance plays a central role. Data provenance is not a topic specific for big data, but it is with big data, distributed, and dynamic databases, where it can be used in its full potentiality. We discuss in this paper some research topics related to privacy and provenance.

The structure of the paper is as follows. In Section 2 we review some of the existing approaches for privacy on standard databases. In Section 3 we focus on the problem for big data. We review its definition and discuss some of the research questions that we consider more relevant with respect to privacy for big data. In Section 4 we focus on the problems related to data provenance. We discuss data provenance and how data provenance interacts with data privacy. The paper finishes with a summary.

2 Data privacy for databases

A large number of mechanisms have been developed for ensuring data privacy. They can be classified according to different dimensions. We classify them [16] according to our knowledge on the type of analysis a third party wants to apply to this data.

- *Data-driven* or general purpose. In this case, we have no knowledge on the type of analysis to be performed by a third party. This is the usual case in which data is published through a server for future use. It also includes the case that data is transferred to a data miner or a data scientist for its analysis as we usually do not know which algorithm will be applied to the data. For this purpose, anonymization methods, also known as masking methods have been developed.
- *Computation-driven* or specific purpose. In this case, we know the exact analysis the third party (or third parties) wants to apply to the data. For example, we know that the data scientist wants to find the parameters of a regression model. This can be seen as the computation of a function or as solving a query for a database without disclosing the database. When a single database is considered and we formulate the problem as answering a query, differential privacy is a suitable privacy model. In the case that multiple databases are considered, the privacy model is based on secure multiparty computation and cryptographic protocols are used for this purpose.
- *Result-driven.* In this case, the analysis (a given data mining algorithm) is also known. The difference with computation-driven approaches is that here we are not worried on the protection of the database per se, but on the protection of some of the outcomes of the algorithm. For example, we know that data scientists will apply association rule mining, and we want to avoid that they infer that people buying diapers also buy beers. Similarly as in computation-driven analysis, prevention of disclosure for this type of analysis is specific to the given computation producing the specific results. In this case, however, the focus is on the knowledge inferred from the data instead of the actual data.

In this paper we focus on anonymization or masking methods. That is, data-driven methods. In short, anonymization algorithms (masking methods) transform a data file X into a file X' with data of less quality. This quality reduction ensures a certain privacy level according to some pre-established privacy model. This is an approach that can be applied to any type of database. It has been successfully applied to, for example, databases, documents, search logs, and social networks.

In addition, the approach is valid not only for protecting data from a syntactic point of view, but also from a semantic point of view. That is, taking into account the meaning of the terms and concepts in the data. For example, when we have words and categories in documents and search logs. For this purpose, we can use masking methods that use ontologies (as e.g. wordnet and ODP) to protect the data.

As masking methods modify the original data reducing its quality, three main research questions appear in the process. The first one is how to reduce the quality of the data. This is done by the masking methods themselves. There is a plethora of methods for this. Then, as data is modified we need to be sure that there is no information loss in the process or that this information loss is as low as possible. In other words, data utility is not reduced substantially in the masking process. Information loss measures are defined to quantify this information loss. Finally, although the quality of the data is reduced to avoid the disclosure of sensitive information, there is no guarantee that all methods satisfy this property. Disclosure risk measures have been defined to quantify the disclosure risk of anonymized data, and they are tightly related to privacy models.

As a summary, we list below the three main research issues related to masking methods.

- Masking methods. Methods that given a database X transform it into another one X' with less quality. Masking methods are usually classified into three categories: perturbative, non-perturbative and synthetic data generators. Perturbative methods reduce the quality by means of modifying the data introducing some kind of error into the data. Noise addition and multiplication, microaggregation and rank swapping are examples of perturbative methods. Non-perturbative methods reduce the quality of the data making them less detailed (but not erroneous). Generalization and suppression are examples of them. Synthetic data generators replace the original data by data generated from a model, which has been extracted from the original database. So, the data in X' is not the original data but artificial data generated from the model.
- Information loss measures. They measure in what extent the transformation of X into X' reduces the utility of the data, and the information that is lost in the process. Information loss measures are typically defined in terms of an analysis f to be performed to the data. Then, given this analysis f and the original and anonymized files X and X', we define information loss as

$$\mathrm{IL}_f(X, X') = \mathit{divergence}(f(X), f(X')).$$

where *divergence* is a function that evaluates how far are $f(X)$ and $f(X')$. A distance on the space of $f(X)$ can be used for this purpose. Naturally, we expect *divergence*$(Y, Y) = 0$ for all Y. Typical examples of functions f include some statistics (means, variances, covariances, regression coefficients), as well as machine learning algorithms (clustering and classification algorithms). Specific measures for some types of databases have also been considered in the literature (e.g., measures on graphs).
- Privacy models and disclosure risk measures. They focus on what extent anonymized (masked) data still contains sensitive information that can be used to compromise the privacy of the individuals of the database.

3 Data privacy for big data

In this section we propose a few open research questions related to big data. To do so, we outline first a definition of big data, and the major difficulties we find with respect to disclosure risk in big data.

3.1 Big data

There exists several definitions for big data based on the characteristics of the data. The well-known definition based on the 3Vs underlines volume, velocity, and variety as the main characteristics of big data. There are other definitions that expand this definition with additional terms. They are the definitions based on 4Vs, 5Vs, or even 7Vs.

- Volume. Databases include huge amounts of data. For example, facebook generated 4 new petabytes of data per day in October 2014 (see [21]).
- Velocity. Data is flowing to the databases in real time: real time streams of data flowing from diverse resources. Either from sensors or from internet (from e-commerce or social media).
- Variety. Data is no longer of a single type (or a few simple types). Databases include data from a vast range of systems and sensors in different formats and datatypes. This may include unstructured text, logs, and videos.

3.2 Moving privacy to big data: disclosure risk

For big data, in principle, the same research questions mentioned in the previous section appear. We need to develop masking methods, information loss measures and disclosure risk measures. For them, we need to take into account that the amounts of data are larger, and thus we need to deal with the corresponding computational problems. Nevertheless, besides of that, a new issue appears: there is a new level for privacy risk. This new level of risk is caused by the following three problems.

- Lack of control and transparency. It is more and more difficult to know who has our data. There are different organizations that can have information about ourselves without us knowing it. Information is gathered from sensors and cameras, obtained through screening posts in social networks, and from analysis of web searches. Note also the case of tracking cookies. Finally, there are data brokers that gather as much information as possible about citizens.
- Linkability. It is usual for big data to link databases to improve the amount and quality of the information. Linking databases increase the risk of identification as there is more information for each individual. Note that the more information we have on individuals the easier to reidentify them, and the more difficult to protect them.
- Inference and data reusability. There exist effective inference algorithms that infer sensitive information (e.g., sexual orientation, political and religious affiliation [12]). One of the main goals of big data analytics is to use existing data for new purposes. This increases the inference ability. As a side effect, data is never deleted waiting for future use.

In the next section we propose a few research lines for data privacy for big data. They are proposed in relation to these three problems just mentioned.

3.3 Open research questions for big data

We propose in this section a few research questions related to big data. The first one is about the need to inform users about the risks of inference due to big data. Then, we propose some lines related to anonymization of *stand-alone* and *linked* databases. It follows another question related to the need of developing (and using) user privacy. We also discuss the need of developing efficient algorithms for data protection in data privacy. This need is both for user privacy, and respondent and holder privacy. The last one is about data provenance, an issue that is further developed in Section 4.

These lines of research are based on our own work (see e.g. [1, 18, 17]), and on the research lines discussed in [15].

- **Issue #1**. Technology should help people to know what others know and can infer about them.
  As we have stated above, effective machine learning and data mining algorithms can infer sensitive information. Some of these models use data that does not seem *a priori* sensitive. It is *insufficient* that we protect sensitive information without protecting what permits us to infer sensitive information. Technology should help people to know about this, and e.g. provide tools in social networks to make people aware of this fact.
- **Issue #2**. Databases should be anonymized/masked in origin. Machine learning algorithms for masked data are required.
  On the one hand, there exist masking methods that are effective in the sense that they achieve low information loss (with loss disclosure risk). On the other hand, there are machine learning and data mining algorithms that are

resitant to errors. In the same direction, not all data is equally important for machine learning algorithms, and some data mining algorithms for big data do not use all data but only a sample of them. Because of that, it is meaningful to consider privacy by design machine learning algorithms. That is, machine learning methods that are appropriate for data that has already been protected.

  Preprocessing methods for machine learning (dimensionality reduction, sampling, etc.) should be combined with and integrated to masking methods. Masking methods can be seen as methods to introduce noise and reduce quality, but they can also be seen as methods for dimensionality reduction. See e.g. the case of microaggregation and, in general, methods to achieve k-anonymity. They reduce the number of (different) records in a database by means of generalization or clustering (i.e., building centroids). These generalized records or centroids can be seen as more consolidated (error-free) data.

- **Issue #3**. Anonymization needs to provide controlled linkability.

  We have reported that linkability is one of the basic components of big data. Companies want to combine databases to increase the information about individuals (enlarging the set of variables/attributes available on them). If databases are anonymized in origin, we need ways to ensure that these databases can still be somehow linked in order to fulfill big data requirements. k-anonymity allows linkability at group level. Algorithms for controlled linkability are needed, as well as methods that can exploit e.g. linkability at group level.

- **Issue #4**. Privacy models need to be composable.

  Given several data sets with a given privacy guarantee, their combination needs to satisfy also the privacy requirements. There are results on the composability of differential privacy. See e.g. [15].

- **Issue #5**. User privacy should be in place: decentralized anonymity.

  User privacy [17] is when users have an active role in ensuring their own privacy. For this purpose, there are methods to protect the identity of the users as well as to protect their data. For example, there are methods for user privacy in communications and in information retrieval.

  While the research questions mentioned above are to be implemented and used by data holders, user privacy provides users with tools to be used by themselves. User privacy permits that data are anonymized before their transmission to data collectors (or to the service provider). So, there is no need to trust the data collector. Local anonymization and collaborative anonymization are keywords for tools for user privacy.

- **Issue #6**. Methods for big data.

  Big data have particularities (the three or more Vs discussed above) that have to be taken into account when developing methods for ensuring privacy. These particularities are for both respondent and holder privacy (i.e., methods applied by data holders) and for user privacy. We can distinguish three types of situations.

    - **Issue #6.1**. Large volumes of data. Efficient algorithms are required for data of high dimension. Algorithms are required for producing masked databases, but also for computing information loss measures and disclosure risk. There exist already some masking methods that have been developed with efficiency in mind for standard databases (e.g., some algorithms for microaggregation), for graphs and social networks (e.g., based on random noise on edges, on generalization and microaggregation), and for location privacy. New methods are needed.
    - **Issue #6.2**. Dynamic data. When data changes with respect to time, we may need to publish several copies of the database. In this case, specific data masking algorithms are required. Note that independent application of algorithms for *k*-anonymity to the same database can cause disclosure. So, the same applies when the database has changed between two applications of the algorithm.
    - **Issue #6.3**. Streaming data. Data is received continuously and should be processed as soon as possible because we cannot hold them and process them later. In this case, difficulties arise because at any time information is only partial. Methods based on sliding windows have been developed for this purpose.
- **Issue #7**. Data provenance and data privacy.
  The new EU General Data Protection Regulation grants citizens the right to rectify and delete the information about themselves in companies. Data provenance are the data structures that permit companies to know where the information of customers and users is in their databases. Different open research questions appear in the crossroad between provenance and privacy. One of them is the fact that data provenance can contain sensitive information and, thus, privacy technologies needs to be applied to it. At the same time, the fact that data can be modified using data provenance according to customers' requirements poses new privacy problems. We discuss these issues in more detail in the next section. These research topics can also be considered for databases of small and medium size but it is with big data that the research becomes challenging.

4 Data privacy and data provenance

Data provenance is becoming a key issue in data management, and can have a great impact in data privacy. Despite its relevance it has not been given much attention until recently from the data privacy community. Information provided as provenance can be used to improve privacy data mechanisms, but it is important to note that provenance itself has to be protected from inferences [14]. In the era of big data and online social networks, data provenance is also useful to help users to assess the validity and trust of the information. For instance, it can help to identify rumormongers and disinformation centers. As we show in Section 4.2 data provenance can play an important role in the future of big data privacy research.

Broadly speaking, data provenance can be seen as metadata or as an annotation of the data. That is, data is expanded with information of the processes that has led to this data. Provenance can be coarse-grained or fine-grained. That is, we can have information on how a bunch of data (i.e., files or databases have been produced) or we can have information particularized at the record or even at the value level. Fine-grained provenance is what makes provenance useful, as it is only in this case that we have detailed information on how any data element has been produced. E.g., we can know who entered the temperature (fever) of a patient, or in which store our client claimed for a discount.

There are different ways to represent data provenance. There are two types of provenance. They are *where provenance* and *why provenance*. *Where provenance* describes the origin of the data, and *why provenance* the process that generated the data. A data element in a database typically proceeds from the combination of previous data elements by means of certain processing functions. Therefore, we need a structure to represent the transformations. The most common approaches are chains [9, 10, 13] and graphs [3, 20].

4.1 Securing data provenance

Secure provenance was introduced to ensure security and privacy to provenance data. Observe that provenance data is sensitive. It may contain information on who and when data was updated. E.g., knowing that a certain doctor has modified data from a patient can lead to disclosure on who is the doctor of whom, what type of illness the patient has, and at what time the patient was at the hospital. Files and databases typically flow within departments and between companies. It is specially important to ensure that these third parties cannot access confidential information contained in the data provenance, whilst allowing them to work with the factual data and update the provenance structure itself. For example, this would allow to perform analysis on the medical data, preserving patient privacy. Hence, provenance data needs to follow these databases and this has to be done ensuring e.g., provenance integrity. Secure data provenance focuses on these type of problems. A few properties have been established as a requirement for secure data provenance [20], [9], and [10]:

- **Distributed.** When databases flow through untrusted environments, and provenance data is associated to them, we need secure data provenance systems to be defined so that they work in a distributed environment. We cannot use a centralized approach with trusted hardware.
- **Integrity.** In distributed environments it is important that nobody can forge provenance data. Provenance data is transmitted and provenance structures are modified to add the new processes applied to the data. Nevertheless, as stated above the structure is immutable and no adversary can be granted to change any part of it. In addition, the provenance system should not allow the modification of a value without expanding the provenance structure. Finally, deletion of provenance data should not cause that a record of the database is unreadable. Additional aspects to be taken into account is to consider

collusions of intruders (that coalitions of intruders should not be able to attack integrity), repudiation (that intruders should not be able to repudiate a record as it was not theirs) or creating forged structures (intruders should not be able to create new provenance structures).

- **Availability.** We are interested in providing security mechanisms to ensure provenance data availability. Auditors should be able to access provenance information in a secure, fast and reliable manner to perform any required operation, e.g. verify the integrity of an ownership sequence without knowing the individual records.
- **Privacy and confidentiality.** We need to ensure that disclosure does not take place, and this is needed for both the database and the provenance data. Only authorized users can access the information.

These properties need to be combined with the two properties that are general for any provenance system. They are, completeness and efficiency:

- **Completeness.** That is, that all actions that are relevant to computation should be detected and represented in the provenance structure. Note that this is not always easy, because some operations as e.g. cut & paste or manual copy can exclude relevant provenance information.
- **Efficiency.** Data provenance introduces an overhead to the data. Fine-grained provenance can double (or more) the size of a database. In addition, operations on the provenance structure need to be efficient because they also introduce an overhead on the computation time.

All these properties are relevant in the context of big data provenance. Big data is often distributed as different information sources can contribute in a computation or in a decision. Therefore, integrity is a basic aspect. We need that provenance structures are not modified at will, and we need to be sure that only permitted operations are applied to them. Availability is then not only a requirement for auditors but also for the subjects from which the data has been extracted. In order that individuals can access and apply the right to delete or rectify a record, they need to be able to know where their data is or if a certain record contains data that has been generated from their own data.

4.2 Considerations about privacy and provenance

When considering big data associated with provenance data, it is important to clearly define the possible scenarios that may arise for data privacy. An accepted classification of possible situations is given in [2] on the basis of what is protected or where do we want to ensure a given degree of privacy (see Table 1):

- *Case 1*: The data are kept private and provenance data are also private. Both need to be protected and their relation has to be preserved.
- *Case 2*: The data itself are not protected but provenance data are private.
- *Case 3*: Data are private, but the provenance data are not protected.

	Data	Provenance
Case 1	Private	Private
Case 2	Non-private	Private
Case 3	Private	Non-private
Case 4	Non-private	Non-private

Table 1. Cases for privacy and data provenance.

- *Case 4*: No privacy protection are applied to neither the data itself nor the provenance data.

Depending on the different purposes, requirements, and nature of the specific data, a given case might apply. Secure data provenance mainly focuses to the case of private provenance when data is distributed (i.e., we need the system to satisfy the requirements discussed in Section 4.1). In the case of centralized private data standard anonymization techniques can be used if we want a single-shot release of this data.

Some of the problems we encounter when data provenance is used depart from standard solutions of data privacy. We discuss a few examples in the next section.

4.3 Example of privacy problem with provenance information

In this section we illustrate an example of a specific problem that can arise in big data privacy when considering provenance information. This problem might occur when individuals request the deletion of their related data from a given dataset, and thus the model obtained from the data needs to be revised. This operation will be performed by means of provenance data allowing the data operator to know exactly which specific data has to be deleted.

To describe this example, we introduce some notation. We will consider a set X (a file or a database) to which we have applied some masking method ρ to obtain a protected set χ. From χ, using a certain algorithm A we extract a piece of knowledge $\mathit{\Gamma}$. For example, A can be an algorithm to extract decision trees, therefore $\mathit{\Gamma}$ is the decision tree inferred from $\rho(X)$.

The set X is modified with modifications μ to obtain a data set X', which protected with ρ will yield χ' and with algorithm A, the piece of knowledge $\mathit{\Gamma}'$. E.g., $\mathit{\Gamma}'$ is a (different) decision tree inferred from $\rho(X')$.

Notation and procedures are represented in Figure 1.

In most cases μ should not be public since it will lead to reidentification of modified records. In front of this scenario some interesting questions might arise.

- An intruder knows $S \subset X$, $\mathit{\Gamma}$, and $\mathit{\Gamma}'$, can this intruder gain knowledge of μ and $S' \subseteq X \setminus S$ with certainty?
- An intruder knows χ and χ', will this intruder be able to determine μ?

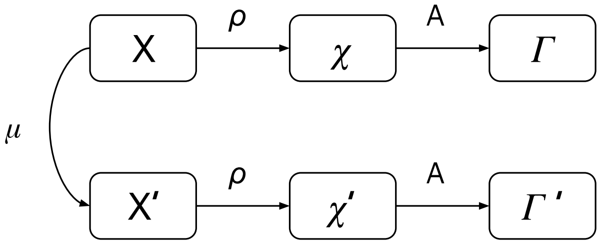

Fig. 1. Example of protected data and its modification

In order to avoid that intruders can make the inferences outlined in the previous lines, privacy models and privacy algorithms can be defined and implemented. In [18] we introduced a privacy model related to the modifications of a database.

5 Summary

In this paper we have proposed a few open questions on the topic of data privacy for big data. On the one hand, we have discussed lines related to stand-alone and linked databases. Among them, we want to stress the need that databases are anonymized in origin, and thus technology is developed to permit controlled linkability and composability.

On the other hand, we have discussed issues related to data provenance, and its relationship with data privacy.

Acknowledgements

This work was supported by Spanish MINECO (project TIN2014-55243-P), and Catalan AGAUR (2014-SGR-691).

References

1. D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.-A., Bourka, A. (2015) Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics, ENISA: European Union Agency for Network and Information Security.
2. Bertino, E., Ghinita, G., Kantarcioglu, M., Nguyen, D., Park, J., Sandhu, R., Sultana, S., Thuraisingham, B., Xu, S., (2014). A roadmap for privacy-enhanced secure data provenance. J Intell Inf Syst 43, 481-501.
3. Braun, U., Shinnar, A., Seltzer, M. (2008) Securing provenance, Proc. HOTSEC 2008.
4. Chaum, D. L. (1981) Untraceable electronic mail, return addresses, and digital pseudonyms, Communications of the ACM 24:2 84-88.

5. Dalneius, T. (1974) The invasion of privacy problem and statistics production – an overview, Statistisk tidskrift 213-225.
6. Dalenius, T. (1977) Towards a methodology for statistical disclosure control, Statistisk Tidskrift 5 429-444.
7. D.E. Denning and J. Schlörer. 1980. A fast procedure for finding a tracker in a statistical database. ACM Trans. Database Syst. 5, 1 (March 1980), 88-102.
8. Duncan, G. T., Elliot, M., Salazar, J. J. (2011) Statistical confidentiality, Springer.
9. Hasan, R., Sion, R., Winslett, M. (2007) Introducing secure provenance: problems and challenges, Proc. StorageSST, ACM, 2007.
10. Hasan, R., Sion, R., Winslett, M. (2009) The case of the fake Picasso: preventing history forgery with secure provenance, Proc. FAST'09.
11. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., de Wolf, P.-P. (2012) Statistical Disclosure Control, Wiley.
12. Kosinski, M., Stillwell, D., Graepel, T. (2013) Private traits and attributes are predictable from digital records of human behavior, PNAS.
13. McDaniel, P., Butler, K., Sion, R., Zadok, E., Winslett, M. (2010) Towards a secure and efficient system for end-to-end provenance, Proc. TAPP 2010.
14. Reuben, J., Martucci, L.A., Fischer-Hübner, S., Packer, H.S., Hedbom, H., Moreau, L., (2016) Privacy Impact Assessment Template for Provenance, 2016 11th International Conference on Availability, Reliability and Security (ARES), pp. 653-660.
15. Soria-Comas, J., Domingo-Ferrer, J. (2016) Big Data Privacy: Challenges to Privacy Principles and Models, Data science and engineering 1:1 21-28.
16. Torra, V., Navarro-Arribas, G. (2014) Data Privacy, WIREs Data Mining and Knowledge Discovery 4:4 269-280.
17. Torra, V. (2017) Data privacy, Springer (forthcoming).
18. Torra, V., Navarro-Arribas, G. (2016) Integral privacy, Proc. CANS 2016. LNCS 10052 661-669.
19. Vaidya, J., Clifton, C. W., Zhu, Y. M. (2006) Privacy Preserving Data Mining, Springer.
20. Zhang, J., Chapman, A., LeFevre, K. (2009) Do you know where your data's been? - Tamper-evident database provenance, Proc. SDM 2009, LNCS 5776 17-32.
21. https://www.brandwatch.com/2016/05/47-facebook-statistics-2016/