Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Бакалавърска програма „Софтуерно инженерство”

Курсов проект

Предмет: Социално-правни аспекти на софтуерното инженерство

Зимен семестър, 2020/2021 год.

СИ-11.01.2021-3-18

Тема №4 : „Anonymize“

Проверил: гл. ас. д-р Калина Георгиева

Изготвили:

Георги Димитров, ф.н. 62256

Деница Дамбова, ф.н. 62364

Добромир Колев, ф.н. 62356

Кристиан Димитров, ф.н.62260

Мирослав Стефанов, ф.н.62282

Съдържание

1. Увод

Преди създаването на Интернет Европа винаги е била моделът за това как нашите данни трябва да бъдат защитени и регулирани. Причината е, че загрижеността на обществото за неприкосновеността на личния живот доминира в бизнес сферата, като гарантира, че винаги се вземат предвид строги правила за това как компаниите използват личните данни на своите граждани.

През април 2016 г. Европейският парламент прие Законът за защита на личните данни, заместващи остарялата си Директива за защита на данните, приета през 1995 г. За разлика от регламент, директива позволява на всеки от двадесет и осемте членове на ЕС да приеме и персонализира закона според нуждите на своите граждани, докато регламент изисква пълното му приемане, без да има свобода на действие от всички 28 страни на второ място. В този случай GDPR изисква всички 28 държави от ЕС да се съобразят.

Проблемът с директивата е, че тя вече не е от значение за днешната цифрова ера. Нейните разпоредби не се отнасят до начина, по който данните се съхраняват, събират и прехвърлят днес - дигитална ера. Подобно на много наредби и закони в целия ЕС и САЩ, тези разпоредби не са в състояние да издържат на темповете на технологичния напредък.

Сега да изясним няколко термина:

*„**Личните данни** са всяка информация, която се отнася до идентифициран или идентифицируем жив индивид. Различни части от информация, събрани заедно, могат да доведат до идентифициране на конкретно лице, също представляват лични данни“.*

Субект на личните данни - *„(...) физическо лице, което може да бъде идентифицирано, пряко или непряко, по-специално чрез идентификатор като име, идентификационен номер, данни за местонахождение, онлайн идентификатор или по един или повече признаци, специфични за физическата, физиологичната, генетичната,*

психическата, умствената, икономическата, културната или социална идентичност на това физическо лице (...)“.

Обработване на личните данни - *всички действия, извършени с ЛД чрез автоматични или не средства (събиране, записване, организиране, структуриране, съхранение, адаптиране или промяна, извличане, консултиране, употреба, разкриване чрез предаване, разпространяване или друг начин, по който данните стават достъпни, подреждане или комбиниране, ограничаване, изтриване или унищожаване).*

Основната нормативна уредба за защита на личните данни, действаща на територията на Република България се състои от Конвенция № 108 на Съвета на Европа Модернизирана версия, ЗЗЛД, ПДКЗЛДНА и Регламент (ЕС) 679/2016 Директива 95/46 (ЕО).

2. Нормативни източници

2.1. Конвенции

Конвенция № 108 на Съвета на Европа Модернизирана версия

Целта на тази конвенция е да гарантира на територията на всяка страна за всяко физическо лице, независимо от неговата националност и местопребиваване, зачитане на неговите права и основни свободи и по конкретно правото му на личен живот по отношение на автоматизираната обработка на лични данни, отнасящи се до него ("защита на данните").

2.2. Регламенти

Регламент (ЕС) 679/2016 Директива 95/46 (ЕО)

Регламент относно защитата на физическите лица във връзка с обработването на лични данни и относно свободното движение на такива данни и за отмяна на Директива 95/46/ЕО (Общ регламент относно защитата на данните)

2.3. Закони

ЗЗЛД (Закон за защита на личните данни)

Този закон урежда обществените отношения, свързани със защитата на правата на физическите лица при обработване на личните им данни.

2.4. Правилници

ПДКЗЛДНА (Правилник за дейността на Комисията за защита на личните данни и на нейната администрация)

Чл.1. С този правилник се уреждат структурата и организацията на работа на Комисията за защита на личните данни, наричана „комисията”, и на нейната администрация

Чл.2. Дейността на комисията се осъществява при спазване принципите на законност, йерархичност при прилагане на нормативните актове, добросъвестност, справедливост, колегиалност, търсене на обективната истина, служебното начало, самостоятелност и безпристрастност, публичност, бързина и процесуална икономия, последователност и предвидимост, равенство на страните в производството.

3. Ненормативни източници

3.1 Техники за постигане:

3.1.1. Рандомизация (Randomization) – фамилия от техники, които променят достоверността на данните с цел премахване на силната връзка между данните и субекта.

3.1.1.1. Добавяне на шум (Noise addition) – техниката на добавяне на шум е полезна, когато атрибутите могат да имат важно неблагоприятен ефект върху субектите и се състои в модифициране на атрибути в набора от данни, така че те са по-малко точни, като същевременно запазват общото разпределение.

3.1.1.2. Пермутация (Permutation) – състои се от разбъркване на стойностите на атрибутите в таблица, така че някои от тях са изкуствено свързани с различни данни. Полезно е когато е важно да се запази точно разпределение на всеки атрибут в рамките на набора от данни.

3.1.1.3. Диференциална поверителност (Differential privacy) – идеята е да се направи случайна малка замяна в данните, така че резултата от единична заявка / опит да не може да определи много точно нещо за индивида.

3.1.2. Генерализация (Generalization) – второто семейство техники за анонимизиране. Този подход се състои от обобщаване или разреждане на атрибутите на данните чрез модифициране на съответната скала или порядък (т.е. регион, а не град, месец а не седмица).

3.1.2.1. Агрегиране и K-анонимност (Aggregation and K-anonymity) – имат за цел да попречат определен субект от данните да бъде идентифициран, като го групират с поне k други субекта. За да се постигне това, стойностите на атрибутите са генерализирани до такава степен, че всеки субект споделя една и съща стойност.

3.1.2.2. L-разнообразие/ T-близост (L-diversity/T-closeness):

- L-разнообразието разширява k-анонимността, за да гарантира, че детерминираните атаки вече не са възможни, уверявайки се, че във всеки клас на еквивалентност всеки атрибут има поне l различни стойности.

- T-близостта е усъвършенстване на l-разнообразието, като цели да създаде еквивалентни класове, които приличат на първоначалното разпределение на атрибутите в таблицата. Тази техника е полезна, когато е важно е да се запазят данните възможно най-близки до оригинала.

3.2. Анонимизация

Анонимизацията на данни е вид дезинфекция на информацията, чието намерение е защита на поверителността. Това е процесът на премахване на лична информация от набори от данни, така че хората, които данните описват, да останат анонимни. Промените в данните са необратими, т.е. оригиналните данни не могат да бъдат възстановени.

За да е успешна техниката на анонимизация, е необходимо да е стабилна въз основа на три критерия:

- възможно ли е да се идентифицира дадено лице,
- възможно ли е да се да се свързват записи, отнасящи се до дадено лице,
- възможно ли е да се изведе информация относно дадено лице?

3.3. Примери за техниките

3.3.1. Добавяне на шум (Noise addition)

При запазване на височината на даден човек да се направи умишлено разминаване с +-10см. По този начин нито лицето може да бъде идентифицирано от външни страни, нито данните могат да бъдат възстановени или да бъде установено как са били променени.

3.3.2. Пермутация (Permutation)

Нека разгледаме множество атрибути от таблица с медицински данни като причина за настаняване в болница, симптоми, отделение, в което пациентът е настанен и т.н. Забелязваме, че има силна логическа връзка между стойностите и ако решим да

разбъркаме само по една стойност на ред, това лесно може да бъде забелязано и дори да бъдат възстановени оригиналните.

3.3.3. Диференциална поверителност (Differential privacy)

В едно социално проучване е зададен въпрос "Притежаваш ли атрибут A?" Хвърля се монета. Ако се падне ези, тогава се мята монетата пак, но независимо от резултата се дава честен отговор. Ако пък се падне тура, монетата се мята пак и ако се падне ези означава отговор "Да", в противен случай - "Не". По този начин, ако означим с p вероятността случаен човек да притежава атрибута, то можем да пресметнем, че очаквания брой "Да" е ((1/4) + p/2)*n, където n е броя опити. По този начин можем да намерим търсеното p без да имаме достоверна информация за конкретните лица.

3.3.4. Агрегиране и K-анонимност (Aggregation and K-anonymity)

Например, вместо да пазим град, да пазим държава. На рождени дати да съпоставяме диапазон от време - месец или година. На конкретни числови данни да съпоставяме интервал. Например, заплата от 1000 до 1500.

3.3.5. L-разнообразия/ T-близост (L-diversity/T-closeness):

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Чувствителните атрибути (Настинка, Херпес, Акне) са разнообразни в рамките на различните класове на еквивалентност.

4. Приложение върху реални данни

Разгледаните методи от т. 3 ще бъдат строго приложени върху регистър на сдруженията на собствениците на територията на община Джебел. Прави впечатление, че правата на гражданите за защита на личните данни не са спазени, поради което представеният образец от регистъра подлежи на частична обработка.

№ по ред	Регистрационен номер	Наименование	Адрес	Предмет на дейност[2]	Срок	Представени идеални части в % от етажната собственост	Членове на управителния съвет	Начин на представителство
1	1/26.02.2015 г.	„Младост 2 - Джебел"	гр.Джебел, ул."Тракия" № 8	"За усвояване на средства от фондовете на Европейския съюз и/или от държавния или общинския бюджет, безвъзмездна помощ и субсидии и/ или използване на собствени средства с цел ремонт и обновяване на сгради в режим на етажна собственост"	безсрочен	82%	Председател: Росен Станилов Добрев	Чрез председател на Управителен съвет

Първа колона за номер по ред на създаване на сдружение се премахва напълно.

От втора колона се отстраняват номер по ред и дата на основаване на сдружение, като се запазва единствено година. Заглавието се преименува от “регистрационен номер” на “регистрационна година”.

Наименование на сдружението, което в случая представлява блок или улица от адреса, се съкращава до първа буква и се отстраняват всякакви конкретизации по адреса.

От адрес на сдружението се отстраняват селище и номер на улица, като нейното наименование се замества със съответен брой символи ‘*’.

От предмет на дейност се изчиства всичко освен цел на сдружението. Тя се обобщава с абревиатура, която се образува от първите букви на всяка дума, участваща в целта на сдружението.

Премахва се колоната, указваща краен срок на сдружение.

Представените идеални части в % от етажната собственост се заменят със символите “***%”.

В колоната с членове на управителен съвет се елиминират всякакъв вид длъжности. Имената на членовете се заменят с инициали на техните имена и фамилии.

Колоната, която съдържа начин на представителство, също се премахва.

След строго прилагане на корекциите за защита на лични данни, примерът от регистъра придобива следния вид:

Регистрационна година	Наименование	Адрес	Предмет на дейност[2]	Представени идеални части в % от етажната собственост	Членове на управителния съвет
2015 г.	„М.“	ул.“*******“	“росрес“	***%	Р. Д.

5. Заключение

Техниките на анонимизация са обект на интензивни изследвания и това постоянно показва, че всяка техника има своите предимства и недостатъци. В повечето случаи не е възможно да се дадат дори минимални препоръки за параметри, които да се използват, тъй като всеки набор от данни трябва да се разглежда като отделен случай.

В много случаи анонимизиран набор от данни все още може да представлява остатъчен риск за субектите на данни. Наистина, дори когато вече не е възможно прецизното извличане на

записа на дадено лице, остава възможно да се събира информация за това лице с помощта на други източници на информация, която е достъпна (публично или не).

Сдруженията усвояват средства от еврофондове и дори и на пръв поглед тези данни да изглеждат безобидни, ако попаднат в ръцете на някой злонамерен тип, посочените лица в таблицата може да станат жертва на изнудване и/или рекет.