GRUPA ROBOCZA ART. 29

0829/14/PL
WP216

Opinia 05/2014 w sprawie technik anonimizacji

przyjęta w dniu 10 kwietnia 2014 r.

GRUPA ROBOCZA DS. OCHRONY OSÓB FIZYCZNYCH W ZAKRESIE PRZETWARZANIA DANYCH OSOBOWYCH

powołana na mocy dyrektywy 95/46/WE Parlamentu Europejskiego i Rady z dnia 24 października 1995 r.,

uwzględniając art. 29 i 30 tej dyrektywy,

uwzględniając swój regulamin wewnętrzny,

PRZYJMUJE NINIEJSZĄ OPINIĘ:

STRESZCZENIE

W niniejszej opinii grupa robocza przeprowadza analizę skuteczności i ograniczeń istniejących technik anonimizacji względem ram prawnych UE w zakresie ochrony danych oraz przedstawia zalecenia dotyczące postępowania w przypadku tych technik poprzez uwzględnienie ryzyka szczątkowego identyfikacji, które jest właściwe dla każdej z tych technik.

Grupa robocza uznaje potencjalną wartość anonimizacji, w szczególności jako strategii czerpania korzyści z „otwartych danych" przez osoby fizyczne i ogół społeczeństwa przy jednoczesnym ograniczaniu ryzyka dla danych osób. Studia przypadków i publikacje badań wykazały jednak, jak trudne jest utworzenie prawdziwie anonimowego zbioru danych przy jednoczesnym zachowaniu tylu podstawowych informacji, ile jest konieczne do realizacji zadania.

W świetle dyrektywy 95/46/WE i innych odpowiednich instrumentów prawnych UE anonimizacja wynika z przetwarzania danych osobowych w celu nieodwracalnego uniemożliwienia identyfikacji. Przy przeprowadzaniu anonimizacji administratorzy danych muszą wziąć pod uwagę szereg elementów, uwzględniając wszystkie sposoby, jakie mogą zostać wykorzystane (przez administratora danych lub jakąkolwiek osobę trzecią) w celu przeprowadzenia identyfikacji.

Anonimizacja stanowi dalsze przetwarzanie danych osobowych; w związku z tym musi spełniać wymóg zgodności poprzez uwzględnienie podstaw prawnych i okoliczności dalszego przetwarzania. Ponadto zanonimizowane dane nie są objęte zakresem przepisów dotyczących ochrony danych, ale osoby, których dane dotyczą, nadal mogą być upoważnione do ochrony na mocy innych przepisów (takich jak te chroniące poufność komunikacji).

W niniejszej opinii opisano główne techniki anonimizacji, mianowicie randomizację i uogólnianie. W szczególności w niniejszej opinii omówiono dodawanie zakłóceń, permutację, prywatność różnicową, agregację, k-anonimizację, l-dywersyfikację i t-bliskość. Wyjaśniono zasady tych technik, ich zalety i wady oraz powszechne błędy i niepowodzenia związane ze stosowaniem każdej z technik.

W niniejszej opinii omówiono niezawodność każdej techniki w oparciu o trzy kryteria:

(i) czy nadal możliwe jest wyodrębnienie konkretnej osoby fizycznej?;

(ii) czy nadal możliwe jest powiązanie zapisów dotyczących konkretnej osoby fizycznej?; oraz

(iii) czy można wywnioskować informacje w odniesieniu do konkretnej osoby fizycznej?

Znajomość głównych zalet i wad każdej techniki pomaga wybrać sposób opracowania odpowiedniego procesu anonimizacji w danym kontekście.

Uwzględniono również pseudonimizację w celu wyjaśnienia niektórych zagrożeń i błędnych przekonań: pseudonimizacja nie jest metodą anonimizacji. Technika ta ogranicza jedynie możliwość powiązania zbioru danych z prawdziwą tożsamością osoby, której dane dotyczą, i w związku z tym jest ona użytecznym środkiem bezpieczeństwa.

W niniejszej opinii uznano, że techniki anonimizacji mogą zapewnić gwarancje prywatności i mogą być wykorzystywane w celu wygenerowania efektywnych procesów anonimizacji, ale wyłącznie wtedy, gdy ich stosowanie jest odpowiednio zaprojektowane – oznacza to, że aby osiągnąć docelową anonimizację przy jednoczesnym wytwarzaniu pewnych użytecznych danych, należy jasno określić konieczne warunki (kontekst) i cel lub cele procesu anonimizacji. O optymalnym rozwiązaniu należy decydować w ramach poszczególnych przypadków, w miarę możliwości poprzez stosowanie połączenia różnych technik, uwzględniając jednocześnie zalecenia praktyczne opracowane w niniejszej opinii.

Ponadto administratorzy powinni uwzględniać, że zanonimizowane zbiory danych mogą nadal stanowić ryzyko szczątkowe dla osoby, której dane dotyczą. W praktyce z jednej strony anonimizacja i ponowna identyfikacja stanowią aktywne dziedziny badań i regularnie publikowane są nowe odkrycia w tym zakresie, z drugiej strony nawet zanonimizowane dane, takie jak statystyki, mogą być wykorzystywane w celu wzbogacenia istniejących profili poszczególnych osób fizycznych, co prowadzi do powstawania nowych problemów związanych z ochroną danych. Nie należy zatem uznawać anonimizacji za działanie jednorazowe, a administratorzy danych powinni regularnie przeprowadzać ponowną ocenę występującego ryzyka.

1 Wprowadzenie

W miarę jak urządzenia, sensory i sieci tworzą duże ilości i nowe rodzaje danych, a koszty przechowywania danych stają się nieistotne, rośnie zainteresowanie społeczeństwa ponownym wykorzystywaniem tych danych oraz rośnie jego zapotrzebowanie na te dane. „Otwarte dane" mogą przynieść społeczeństwu, osobom fizycznym i organizacjom wyraźne korzyści, ale jedynie wtedy, gdy przestrzegane są prawa wszystkich osób do ochrony ich danych osobowych i życia prywatnego.

Anonimizacja może być dobrą strategią zachowania korzyści i ograniczenia ryzyka. Gdy zbiór danych jest już faktycznie zanonimizowany i nie ma już możliwości zidentyfikowania poszczególnych osób fizycznych, prawo UE o ochronie danych nie ma dalej zastosowania. Ze studiów przypadków i publikacji badań jasno jednak wynika, że utworzenie prawdziwie anonimowego zbioru danych na podstawie bogatego zbioru danych osobowych przy jednoczesnym zachowaniu odpowiedniej ilości informacji podstawowych niezbędnych na potrzeby wykonania zadania nie jest propozycją łatwą do zrealizowania. Na przykład zbiór danych uważany za anonimowy może być połączony z innym zbiorem danych w taki sposób, że istnieje możliwość zidentyfikowania co najmniej jednej osoby fizycznej.

W niniejszej opinii grupa robocza przeprowadza analizę skuteczności i ograniczeń istniejących technik anonimizacji względem ram prawnych UE w zakresie ochrony danych oraz przedstawia zalecania dotyczące ostrożnego i odpowiedzialnego stosowania tych technik w celu przygotowania procesu anonimizacji.

2 Definicje i analiza prawna

2.1. Definicje w kontekście prawnym UE

W motywie 26 dyrektywy 95/46/WE odniesiono się do anonimizacji w celu wykluczenia zanonimizowanych danych z zakresu przepisów dotyczących ochrony danych:

> „Zasady ochrony danych muszą odnosić się do wszelkich informacji dotyczących zidentyfikowanych lub możliwych do zidentyfikowania osób; w celu ustalenia, czy daną osobę można zidentyfikować, należy wziąć pod uwagę wszystkie sposoby, jakimi może posłużyć się administrator danych lub inna osoba w celu zidentyfikowania owej osoby; zasady ochrony danych nie mają zastosowania do danych, którym nadano anonimowy charakter w taki sposób, że podmiot danych nie będzie mógł być zidentyfikowany; zasady postępowania w rozumieniu art. 27 mogą być przydatnym instrumentem w udzielaniu wskazówek co do sposobów nadawania danym charakteru anonimowego oraz zachowania w formie, w której identyfikacja osoby, której dane dotyczą, nie jest dłużej możliwa"[1].

Z wnikliwej lektury motywu 26 wynika definicja pojęciowa terminu anonimizacja. W motywie 26 wskazano, że w celu zanonimizowania jakichkolwiek danych, dane te muszą być

[1] Ponadto należy zauważyć, że przedmiotowe podejście jest również stosowane w projekcie rozporządzenia UE w sprawie ochrony danych na mocy motywu 23: „Aby ustalić, czy można zidentyfikować daną osobę fizyczną, należy wziąć pod uwagę wszystkie sposoby, jakimi mogą posłużyć się administrator lub inna osoba w celu zidentyfikowania tej osoby".

pozbawione wystarczającej liczby elementów, tak aby nie było już możliwości zidentyfikowania osoby, której dane dotyczą. Dokładniej mówiąc, wspomniane dane należy przetwarzać w taki sposób, aby nie istniała już możliwość wykorzystania ich do zidentyfikowania osoby fizycznej za pomocą „wszystkich sposobów, jakimi może posłużyć się" administrator danych lub osoba trzecia. Istotnym czynnikiem jest fakt, że przetwarzanie musi być nieodwracalne. W dyrektywie nie wyjaśniono, w jaki sposób należy lub można przeprowadzić proces anonimizacji[2]. Nacisk kładzie się na wynik: powstałe dane powinny być takie, aby nie umożliwiały zidentyfikowania osoby, której dane dotyczą, za pomocą „wszystkich" sposobów, jakimi „można" „się posłużyć". W dyrektywie odniesiono się do kodeksów postępowania jako do narzędzia określania możliwych mechanizmów anonimizacji oraz zachowania danych w formie, która sprawia, że identyfikacja osoby, której dane dotyczą, „nie jest dłużej możliwa". Zatem dyrektywa wyraźnie ustanawia bardzo wysoki standard.

W dyrektywie o prywatności i łączności elektronicznej (dyrektywie 2002/58/WE) również w bardzo podobny sposób odniesiono się do „anonimizacji" i „danych anonimowych". Motyw 26 stanowi, że:

> „Dane dotyczące ruchu wykorzystywane w marketingu usług komunikacyjnych lub dostarczenia usług tworzących wartość dodaną powinny również zostać usunięte lub uczynione anonimowymi po dostarczeniu usług".

Podobnie art. 6 ust. 1 stanowi, że:

> „Dane o ruchu dotyczące abonentów i użytkowników przetwarzane i przechowywane przez dostawcę publicznej sieci łączności lub publicznie dostępnych usług łączności elektronicznej muszą zostać usunięte lub uczynione anonimowymi, gdy nie są już potrzebne do celów transmisji komunikatu, bez uszczerbku dla przepisów ust. 2, 3 i 5 niniejszego artykułu oraz art. 15 ust. 1".

Co więcej, zgodnie z art. 9 ust. 1:

> „W przypadku gdy dane dotyczące lokalizacji inne niż dane o ruchu, odnoszące się do użytkowników lub abonentów publicznych sieci łączności lub publicznie dostępnych usług łączności elektronicznej, mogą być przetwarzane, przetwarzanie może mieć miejsce tylko wówczas gdy dane te są anonimowe, lub za zgodą użytkowników lub abonentów, w zakresie i przez okres niezbędny do świadczenia usługi tworzącej wartość dodaną".

Podstawową przesłanką jest to, że wynik anonimizacji jako techniki stosowanej do danych osobowych powinien być, przy obecnym stanie technologii, zarówno trwały, jak i możliwy do usunięcia, tj. uniemożliwiający przetwarzanie danych osobowych[3].

---

[2] Pojęcie to omówiono dalej na s. 8 niniejszej opinii.

[3] W tym miejscu należy przypomnieć, że anonimizację określa się również w standardach międzynarodowych, takich jak ISO 29100, gdzie jest ona „procesem, w którym informacje umożliwiające identyfikację osoby są nieodwracalnie zmienione w taki sposób, aby nie istniała już możliwość bezpośredniego lub pośredniego zidentyfikowania podmiotu informacji umożliwiających identyfikację osoby przez administratora informacji umożliwiających identyfikację osoby działającego samodzielnie lub we współpracy z jakąkolwiek inną stroną" (ISO 29100:2011). Nieodwracalność zmiany, której poddane zostały dane osobowe w celu umożliwienia bezpośredniej lub pośredniej identyfikacji, jest istotna również dla ISO. Z tego punktu widzenia występuje znaczna zbieżność między zasadami i pojęciami leżącymi u podstaw dyrektywy 95/46. Dotyczy to również definicji, które można znaleźć w niektórych przepisach krajowych (np. we Włoszech, w Niemczech i w Słowenii), w których nacisk położono na brak możliwości identyfikacji i odniesiono się do „nieproporcjonalnych wysiłków" na rzecz ponownej identyfikacji (D, SI). Francuskie prawo o ochronie danych stanowi jednak, że dane pozostają danymi osobowymi, nawet jeżeli ponowna identyfikacja osoby, której dane dotyczą, jest bardzo trudna i mało prawdopodobna, tj. brak jakiegokolwiek przepisu odnoszącego się do testu „zasadności".

2.2. Analiza prawna

Analiza sformułowań odnoszących się do anonimizacji w głównych instrumentach UE dotyczących ochrony danych umożliwia wyróżnienie czterech cech podstawowych:

- anonimizacja może być wynikiem przetwarzania danych osobowych w celu nieodwracalnego uniemożliwienia zidentyfikowania osoby, której dane dotyczą;

- ponieważ w przepisach UE nie ma żadnych norm nakazowych, możliwe jest stosowanie szeregu technik anonimizacji;

- należy nadać znaczenie elementom kontekstowym: należy uwzględnić „wszystkie" sposoby, „jakimi może posłużyć się" administrator danych i osoby trzecie w celu przeprowadzenia identyfikacji, zwracając szczególną uwagę na to, jakiego znaczenia, przy obecnym stanie technologii, nabrało ostatnio sformułowanie „jakimi może posłużyć się" (uwzględniając moc obliczeniową i dostępne narzędzia);

- czynnik ryzyka jest nieodłącznym elementem anonimizacji: należy uwzględnić wspomniany czynnik ryzyka przy ocenianiu ważności każdej techniki anonimizacji – w tym możliwych zastosowań wszelkich danych, które zostały „zanonimizowane" za pomocą takiej techniki – należy również ocenić wagę i prawdopodobieństwo takiego ryzyka.

W niniejszej opinii używa się sformułowania „technika anonimizacji" zamiast terminu „anonimowość" lub „dane anonimowe", aby podkreślić ryzyko szczątkowe właściwe dla ponownej identyfikacji w związku z każdym środkiem techniczno-organizacyjnym mającym na celu zachowanie „anonimowości" danych.

2.2.1. Zgodność z prawem i proces anonimizacji

Po pierwsze, anonimizacja jest techniką stosowaną do danych osobowych w celu uzyskania nieodwracalnej anonimizacji. Dlatego też założeniem wyjściowym jest to, że dane osobowe musiały zostać zgromadzone i przetworzone zgodnie z mającymi zastosowanie przepisami dotyczącymi utrzymywania danych w formacie uniemożliwiającym identyfikację.

W tym kontekście proces anonimizacji, oznaczający przetwarzanie takich danych osobowych w celu osiągnięcia ich anonimizacji, stanowi przykład „dalszego przetwarzania". Przetwarzanie musi zatem spełniać test zgodności na podstawie wytycznych przedstawionych przez grupę roboczą w jej opinii 03/2013 w sprawie ograniczania celu[4].

Oznacza to, że zasadniczo podstawę prawną anonimizacji można znaleźć w każdej z podstaw wymienionych w art. 7 (w tym w uzasadnionym interesie administratora danych), pod warunkiem że spełnione są także wymogi dotyczące jakości danych określone w art. 6 dyrektywy z należytym uwzględnieniem szczególnych okoliczności i wszystkich czynników wymienionych w opinii grupy roboczej w sprawie ograniczania celu[5].

[4] Opinia 03/2013 Grupy Roboczej Art. 29, dostępna pod adresem: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] Oznacza to w szczególności, że ocenę merytoryczną należy przeprowadzić w świetle wszystkich odpowiednich okoliczności ze szczególnym uwzględnieniem następujących najważniejszych czynników:
a) związku między celami, do których gromadzono dane osobowe, a celami ich dalszego przetwarzania;
b) kontekstu, w jakim dane osobowe były gromadzone, i uzasadnionych oczekiwań ze strony osób, których dane dotyczą, co do dalszego wykorzystywania tych danych;
c) charakteru danych osobowych i wpływu dalszego przetwarzania na osoby, których dane dotyczą;

Z drugiej strony należy zwrócić uwagę na przepisy zawarte w art. 6 ust. 1 lit. e) dyrektywy 95/46/WE (ale również przepisy zawarte w art. 6 ust. 1 i art. 9 ust. 1. dyrektywy o prywatności i łączności elektronicznej), ponieważ wskazują one na potrzebę przechowywania danych osobowych „w formie umożliwiającej identyfikację" przez czas nie dłuższy niż jest to konieczne do celów, dla których dane zostały zgromadzone lub dla których są dalej przetwarzane.

Przepis ten sam w sobie silnie podkreśla, że dane osobowe powinny być przynajmniej „domyślnie" anonimizowane (na podstawie różnych wymogów prawnych, takich jak te, o których mowa w dyrektywie o prywatności i łączności elektronicznej w odniesieniu do danych o ruchu). Jeżeli administrator danych chce zatrzymać takie dane osobowe po osiągnięciu celów pierwotnego lub dalszego przetwarzania, należy zastosować techniki anonimizacji w celu nieodwracalnego uniemożliwienia identyfikacji.

Dlatego też grupa robocza stwierdza, że anonimizację jako przykład dalszego przetwarzania danych osobowych można uznać za zgodną z pierwotnymi celami przetwarzania, ale wyłącznie pod warunkiem, że proces anonimizacji zapewnia uzyskanie dokładnie zanonimizowanych informacji w znaczeniu opisanym w niniejszym dokumencie.

Należy również podkreślić, że anonimizacja musi być przeprowadzona zgodnie z ograniczeniami prawnymi przywołanymi przez Trybunał Sprawiedliwości w orzeczeniu w sprawie C-553/07 (College van burgemeester en wethouders van Rotterdam przeciwko v M.E.E. Rijkeboer), odnoszącymi się do potrzeby zachowania danych w formie pozwalającej na identyfikację w celu umożliwienia osobie, której dane dotyczą, na przykład, korzystania z praw dostępu. Trybunał Sprawiedliwości orzekł, że: „Artykuł 12 lit. a) dyrektywy [95/46] nakłada na państwa członkowskie obowiązek ustanowienia prawa dostępu do informacji o odbiorcach lub kategorii odbiorców dotyczących jej danych osobowych oraz treści przekazanych danych nie tylko w odniesieniu do teraźniejszości, lecz również w odniesieniu do przeszłości. Do państw członkowskich należy określenie okresu przechowywania tej informacji oraz odpowiedniego dostępu do tej informacji, który stanowiłby rezultat właściwego wyważenia między, z jednej strony, interesem osoby, której dane dotyczą, w ochronie jej życia prywatnego, w szczególności za pośrednictwem prawa interwencji oraz prawa do wniesienia środka prawnego przewidzianych przez dyrektywę 95/46, a z drugiej strony – obciążeniem, jakie obowiązek przechowywania tej informacji reprezentuje dla administratora danych".

Jest to szczególnie istotne, w przypadku gdy administrator danych opiera się na art. 7 lit. f) dyrektywy 95/46 w odniesieniu do anonimizacji: zawsze należy równoważyć uzasadniony interes administratora danych z prawami i podstawowymi wolnościami osób, których dane dotyczą.

Przykładowo dochodzenie przeprowadzone przez niderlandzki organ ds. ochrony danych osobowych w latach 2012–2013 w sprawie wykorzystywania technologii głębokiej inspekcji pakietów (DPI) przez czterech operatorów sieci ruchomej wykazało podstawę prawną na mocy art. 7 lit. f) dyrektywy 95/46 w odniesieniu do zanonimizowania treści danych o ruchu możliwie najszybciej po zgromadzeniu tych danych. W praktyce art. 6 dyrektywy o prywatności i łączności elektronicznej stanowi, że dane o ruchu dotyczące abonentów i użytkowników przetwarzane i przechowywane przez dostawcę publicznej sieci łączności lub publicznie dostępnych usług łączności elektronicznej należy możliwie najszybciej usunąć lub

d) gwarancji przyjętych przez administratora danych w celu zapewniania uczciwego przetwarzania i uniemożliwienia wszelkiego nieodpowiedniego wpływu na osoby, których dane dotyczą.

uczynić anonimowymi. W analizowanym przypadku, ponieważ jest to dopuszczone na mocy art. 6 dyrektywy o prywatności i łączności elektronicznej, istnieje odpowiednia podstawa prawna w art. 7 dyrektywy o ochronie danych. Można to również przedstawić w odwrotny sposób: jeżeli rodzaj przetwarzania danych nie jest dozwolony na mocy art. 6 dyrektywy o prywatności i łączności elektronicznej, nie może istnieć podstawa prawna w art. 7 dyrektywy o ochronie danych.

2.2.2. Możliwość potencjalnego zidentyfikowania zanonimizowanych danych

Grupa robocza szczegółowo omówiła pojęcie danych osobowych w opinii 4/2007 w sprawie danych osobowych, uwzględniając w szczególności elementy tworzące definicję przedstawioną w art. 2 lit. a) dyrektywy 95/46/WE, w tym część tej definicji, jaką jest sformułowanie „zidentyfikowanej lub możliwej do zidentyfikowania". W tym kontekście grupa robocza stwierdziła także, że „dane, którym nadano anonimowy charakter są danymi anonimowymi, które wcześniej dotyczyły osoby możliwej do zidentyfikowania, lecz której zidentyfikowanie nie jest już możliwe".

W związku z tym grupa robocza wyjaśniła już, że w dyrektywie proponuje się przeprowadzenie testu „sposobów, jakimi można się posłużyć" jako kryterium, które należy zastosować w celu ocenienia, czy proces anonimizacji jest wystarczająco dokładny, tj. czy identyfikacja stała się praktycznie niemożliwa. Szczególny kontekst i okoliczności określonego przypadku mają bezpośredni wpływ na możliwość identyfikacji. W załączniku technicznym do niniejszej opinii przedstawiono analizę wpływu wyboru najbardziej odpowiedniej techniki.

Jak już podkreślono, badania, narzędzia i moc obliczeniowa ulegają zmianom. Dlatego też przedstawienie wyczerpującego wyliczenia okoliczności, w których identyfikacja nie jest już możliwa, nie jest wykonalne ani użyteczne. Niektóre istotne czynniki zasługują jednak na uwzględnienie i omówienie.

Po pierwsze, można stwierdzić, że administratorzy danych powinni w szczególności koncentrować się na konkretnych sposobach, jakie byłyby niezbędne w celu odwrócenia techniki anonimizacji, w szczególności pod względem kosztu i know-how koniecznych do realizacji tych sposobów oraz oceny ich prawdopodobieństwa i wagi. Administratorzy danych powinni na przykład zrównoważyć swoje wysiłki na rzecz anonimizacji i ponoszone koszty (pod względem zarówno wymaganego czasu, jak i wymaganych zasobów) z coraz większą dostępnością środków technicznych umożliwiających zidentyfikowanie osób w zbiorach danych po niskich kosztach, coraz większą powszechną dostępnością innych zbiorów danych (takich jak te udostępnione w związku z polityką „otwartych danych") i z wieloma przykładami niepełnej anonimizacji wiążącej się z późniejszymi szkodliwymi, czasami niemożliwymi do naprawienia skutkami dla osób, których dane dotyczą[6]. Należy zauważyć, że ryzyko identyfikacji może z czasem się zwiększyć i że zależy ono również od rozwoju technologii informacji i komunikacji. Dlatego też regulacje prawne, jeżeli istnieją, muszą być

[6] Co ciekawe, w przedłożonych ostatnio (dnia 21 października 2013 r.) poprawkach Parlamentu Europejskiego do projektu ogólnego rozporządzenia o ochronie danych w szczególności określa się w motywie 23, że „aby ustalić, czy dany sposób może z racjonalnym prawdopodobieństwem posłużyć do zidentyfikowania danej osoby, należy wziąć pod uwagę wszelkie obiektywne czynniki, takie jak koszt i czas potrzebne do zidentyfikowania danej osoby, oraz uwzględnić zarówno technologię dostępną w momencie przetwarzania danych, jak i postęp technologiczny".

sformułowane w sposób neutralny pod względem technologicznym i powinny uwzględniać zmiany w potencjale rozwojowym technologii informacyjnej[7].

Po drugie, „sposoby, jakimi można się posłużyć w celu ustalenia, czy daną osobę można zidentyfikować" są sposobami, jakimi posługuje się „administrator danych lub inna osoba". Konieczne jest zatem zrozumienie, że jeżeli administrator danych nie usunie pierwotnych (umożliwiających identyfikację) danych na poziomie zdarzenia, a przekaże część tego zbioru danych (na przykład po usunięciu lub ukryciu danych umożliwiających identyfikację), powstały zbiór danych nadal stanowi zbiór danych osobowych. Powstały zbiór danych można zakwalifikować jako anonimowy jedynie w przypadku, gdy administrator danych zagregowałby dane, osiągając poziom, na którym nie istnieje już możliwość zidentyfikowania poszczególnych wydarzeń. Na przykład: jeżeli organizacja gromadzi dane dotyczące przemieszczania się osoby fizycznej, wzorce podróżowania tej osoby na poziomie zdarzenia nadal kwalifikowałyby się jako dane osobowe w odniesieniu do każdej strony, o ile administrator danych (lub jakakolwiek inna strona) nadal ma dostęp do oryginalnych danych pierwotnych, nawet jeżeli ze zbioru udostępnionego osobom trzecim usunięto elementy umożliwiające bezpośrednią identyfikację. Jeżeli jednak administrator danych usunąłby dane pierwotne i dostarczył osobom trzecim tylko statystki zagregowane na wysokim poziomie, takim jak „w poniedziałki trasą X jeździ o 160 % więcej pasażerów niż we wtorki", statystyki te kwalifikowałyby się jako dane anonimowe.

Skuteczne rozwiązanie w zakresie anonimizacji uniemożliwia wszystkim stronom wyodrębnienie konkretnej osoby fizycznej ze zbioru danych, tworzenie powiązań między dwoma zapisami w zbiorze danych (lub między dwoma oddzielnymi zbiorami danych) i wnioskowanie jakichkolwiek informacji z tego zbioru danych. Zatem, ogólnie rzecz biorąc, samo usunięcie elementów umożliwiających bezpośrednią identyfikację nie wystarcza do zapewnienia, aby zidentyfikowanie osoby, której dane dotyczą, nie było już możliwe. Często konieczne będzie podjęcie dodatkowych środków w celu zapobieżenia identyfikacji, które to środki również zależą od kontekstu i celów przetwarzania, do jakich przeznaczone są zanonimizowane dane.

PRZYKŁAD:
Profile danych genetycznych stanowią przykład danych osobowych, w przypadku których może występować ryzyko identyfikacji, jeżeli jedyną zastosowaną techniką jest usunięcie tożsamości dawcy, ze względu na niepowtarzalny charakter określonych profili. W literaturze wykazano już[8], że połączenie dostępnych publicznie zasobów genetycznych (np. rejestrów genealogicznych, nekrologów, wyników zapytań wprowadzanych do wyszukiwarek internetowych) z metadanymi dotyczącymi dawców DNA (czasu przekazania, wieku dawcy, jego miejsca zamieszkania) może zdradzać tożsamość określonych osób, nawet jeżeli DNA przekazane zostało „anonimowo".

W obu rodzinach technik anonimizacji – randomizacji danych i ich uogólnianiu[9] – występują niedoskonałości; każda z tych technik może być jednak odpowiednia w określonych okolicznościach i w określonym kontekście do osiągnięcia zamierzonego celu, nie zagrażając prywatności osób, których dane dotyczą. Należy jasno stwierdzić, że „identyfikacja" nie oznacza tylko możliwości uzyskania nazwiska jakiejś osoby lub jej adresu, ale obejmuje również potencjalną możliwość identyfikacji przez wyodrębnienie, tworzenie powiązań i

[7] Zob. opinia 4/2007 Grupy Roboczej Art. 29, s. 15.
[8] Zob. John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, tom 339, nr 6117 (18 stycznia 2013 r.), s. 262.
[9] Główne cechy tych dwóch technik anonimizacji i najważniejsze różnice między nimi opisano w sekcji 3 poniżej („Analiza techniczna").

wnioskowanie. Co więcej, aby miało zastosowanie prawo o ochronie danych, intencje administratora danych lub odbiorcy danych nie mają znaczenia. Dopóki dane umożliwiają identyfikację, mają zastosowanie przepisy w zakresie ochrony danych.

Osoba trzecia może zgodnie z prawem przetwarzać zbiór danych, względem którego zastosowano technikę anonimizacji, (zanonimizowany i udostępniony przez administratora danych pierwotnych) bez potrzeby brania pod uwagę wymogów w zakresie ochrony danych, pod warunkiem że osoba ta nie może (bezpośrednio lub pośrednio) zidentyfikować osób, których dane dotyczą, w pierwotnym zbiorze danych. Osoby trzecie zobowiązane są jednak do uwzględnienia wszelkich czynników kontekstowych i okolicznościowych, o których mowa powyżej (w tym szczególnych cech technik anonimizacji stosowanych przez administratora danych pierwotnych), przy podejmowaniu decyzji dotyczącej sposobu wykorzystywania i w szczególności łączenia takich zanonimizowanych danych do ich własnych celów, ponieważ powstałe konsekwencje mogą wiązać się z różnymi rodzajami ponoszonej przez nie odpowiedzialności. Jeżeli te czynniki i cechy pociągają za sobą niedopuszczalne ryzyko zidentyfikowania osób, których dane dotyczą, przetwarzanie będzie ponownie objęte zakresem prawa o ochronie danych.

Powyższa lista nie ma charakteru wyczerpującego, a jej celem jest raczej przedstawienie ogólnych wytycznych dotyczących podejścia do oceny potencjału w zakresie możliwości identyfikacji, jaki ma określony zbiór danych poddany anonimizacji przy zastosowaniu różnych dostępnych technik. Wszystkie powyższe czynniki można uznać za liczne czynniki ryzyka, które muszą zostać wzięte pod uwagę zarówno przez administratorów danych przy anonimizacji zbiorów danych, jak i osoby trzecie przy wykorzystywaniu tych „anonimowych" zbiorów danych do własnych celów.

2.2.3. Ryzyko związane z wykorzystywaniem zanonimizowanych danych

Rozważając zastosowanie technik anonimizacji, administratorzy danych muszą wziąć pod uwagę następujące zagrożenia:

- szczególnym ryzykiem jest uznawanie danych opatrzonych pseudonimem za równoważne ze zanonimizowanymi danymi. W sekcji dotyczącej analizy technicznej zostanie wyjaśnione, że nie można zrównać danych pseudonimicznych ze zanonimizowanymi informacjami, ponieważ te pierwsze nadal umożliwiają wyodrębnienie konkretnej osoby fizycznej, której dane dotyczą, i tworzenie w odniesieniu do tej osoby powiązań między różnymi zbiorami danych. Istnieje prawdopodobieństwo, że w przypadku psuedonimowości możliwa będzie identyfikacja i dlatego technika ta jest objęta zakresem systemu prawnego dotyczącego ochrony danych. Jest to szczególnie istotne w kontekście badań naukowych, statystycznych lub historycznych[10];

PRZYKŁAD:
Typowym przykładem błędnych przekonań związanych z pseudonimizacją jest głośna sprawa „wycieku danych z AOL (America On Line)". W 2006 r. udostępniono publicznie bazę danych zawierającą dwadzieścia milionów słów kluczy wpisywanych do wyszukiwarki przez ponad 650 000 użytkowników przez okres 3 miesięcy, a jedynym środkiem zachowania prywatności było zastąpienie identyfikatorów użytkowników AOL przypisywanymi numerami. Doprowadziło to do publicznego zidentyfikowania i zlokalizowania niektórych użytkowników. Pseudonimiczne szeregi zapytań wprowadzane do wyszukiwarek internetowych, zwłaszcza w połączeniu z innymi atrybutami, takimi jak adresy IP lub inne parametry konfiguracyjne klienta, mają bardzo dużą moc identyfikacji.

[10] Zob. również opinia 4/2007 Grupy Roboczej Art. 29, s. 18–20.

- drugim błędem jest uznawanie, że właściwie zanonimizowane dane (spełniające wszystkie warunki i kryteria, o których mowa powyżej, i z definicji nieobjęte zakresem dyrektywy o ochronie danych) pozbawiają osoby fizyczne wszelkich zabezpieczeń, przede wszystkim dlatego, że do wykorzystywania tych danych mogą mieć zastosowanie inne przepisy. Na przykład art. 5 ust. 3 dyrektywy o prywatności i łączności elektronicznej uniemożliwia przechowywanie „informacji" każdego rodzaju (w tym informacji innych niż dane osobowe) na urządzeniu końcowym i dostęp do tych informacji bez zgody abonenta/użytkownika, ponieważ jest to część szerszej zasady poufności komunikacji;

- trzecie uchybienie również wynika z nieuwzględnienia wpływu, jaki w określonych okolicznościach mogą mieć właściwie zanonimizowane dane na osoby fizyczne, w szczególności w przypadku profilowania. Ochronę sfery życia prywatnego danej osoby fizycznej zapewniono w art. 8 EKPC i art. 7 Karty praw podstawowych Unii Europejskiej; w związku z tym, mimo że przepisy w zakresie ochrony danych mogą nie mieć już zastosowania do tego rodzaju danych, wykorzystywanie zbiorów danych, które zostały zanonimizowane i udostępnione do wykorzystywania przez osoby trzecie, może skutkować utratą prywatności Wymagane jest zachowanie szczególnej ostrożności w postępowaniu ze zanonimizowanymi informacjami, zwłaszcza w każdym przypadku, gdy takie informacje wykorzystuje się (często w połączeniu z innymi danymi) do celów podejmowania decyzji, które mają wpływ (choćby pośrednio) na poszczególne osoby fizyczne. Jak już podkreślono w niniejszej opinii i jak zostało już sprecyzowane przez grupę roboczą, w szczególności w opinii w sprawie pojęcia „związania celem" (opinia 03/2013)[11], uzasadnione oczekiwania osób, których dane dotyczą, w odniesieniu do dalszego przetwarzania ich danych powinny zostać ocenione w świetle odpowiednich czynników związanych z kontekstem, takich jak charakter stosunku między osobami, których dane dotyczą, a administratorami danych, mające zastosowanie obowiązki prawne, przejrzystość i operacje przetwarzania.

3 Analiza techniczna, niezawodność technologii i typowe błędy

Istnieją różne praktyki i techniki anonimizacji o różnych stopniach dokładności. W niniejszej sekcji odniesiono się do najważniejszych punktów, które administratorzy danych muszą rozważyć przy stosowaniu tych praktyk i technik, poprzez zwrócenie uwagi w szczególności na gwarancję, jaką można osiągnąć dzięki danej technice, mając na względzie aktualny stan technologii i uwzględniając trzy czynniki ryzyka właściwe dla anonimizacji:

- *wyodrębnienie,* które oznacza możliwość wydzielenia niektórych lub wszystkich zapisów identyfikujących określoną osobę fizyczną w zbiorze danych;
- *możliwość tworzenia powiązań*, czyli zdolność do powiązania co najmniej dwóch zapisów dotyczących jednej osoby lub grupy osób, których dane dotyczą (w tej samej bazie danych lub w dwóch różnych bazach danych). Jeżeli atakujący może ustalić (np. w drodze analizy korelacji), że dwa zapisy przypisane są tej samej grupie osób fizycznych, ale nie może wyodrębnić poszczególnych osób w tej grupie, dana technika zapewnia ochronę przed „wyodrębnieniem", ale nie przed możliwością tworzenia powiązań;
- *wnioskowanie*, czyli możliwość wydedukowania ze znacznym prawdopodobieństwem wartości danego atrybutu z wartości zbioru innych atrybutów.

[11] Dostępne na stronie http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

Dlatego też rozwiązanie przeciwdziałające tym trzem czynnikom ryzyka byłoby skuteczne pod względem uniemożliwienia ponownej identyfikacji przeprowadzanej w drodze najbardziej prawdopodobnych i uzasadnionych sposobów, jakie mogą zostać wykorzystane przez administratora danych i każdą osobę trzecią. Grupa robocza podkreśla w związku z tym, że techniki depersonalizacji danych i anonimizacji podlegają prowadzonemu obecnie badaniu i że badanie to wykazało spójnie, iż żadna technika sama w sobie nie jest pozbawiona niedoskonałości. Ogólnie mówiąc, istnieją dwa różne podejścia do anonimizacji: pierwsze opiera się na ***randomizacji***, natomiast drugie opiera się na ***uogólnianiu***. Niniejsza opinia dotyczy również innych pojęć, takich jak *pseudonimizacja, prywatność różnicowa, l-dywersyfikacja, t-bliskość.*

W przedmiotowej sekcji niniejszej opinii użyto następującej terminologii: zbiór danych składa się z różnych zapisów dotyczących określonych osób fizycznych (osób, których dane dotyczą). Każdy zapis związany jest z jedną osobą, której dane dotyczą, i składa się ze zbioru wartości (lub „wpisów", np. 2013) w odniesieniu do każdego atrybutu (np. roku). Zbiór danych jest zbiorem zapisów, które można alternatywnie przedstawić w formie tabeli (lub zbioru tabel) lub w formie opisanego/ważonego grafu, co obecnie jest coraz częściej stosowane. Przykłady przedstawione w niniejszej opinii będą się odnosiły do tabel, ale mają one również zastosowanie do innej formy graficznego przedstawienia zapisów. Połączenia atrybutów dotyczących osoby lub grupy osób, których dane dotyczą, mogą być określane jako *quasi*-identyfikatory. W niektórych przypadkach zbiór danych może zawierać wiele zapisów dotyczących tej samej osoby fizycznej. „Atakujący" jest osobą trzecią (tj. nie jest ani administratorem danych, ani przetwarzającym), która przypadkowo lub specjalnie przeprowadza ocenę zapisów pierwotnych.

3.1. Randomizacja

Randomizacja tworzy rodzinę technik, która zmienia prawdziwość danych w celu wyeliminowania ścisłego związku między danymi a konkretną osobą fizyczną. Jeżeli dane charakteryzują się wystarczającą niepewnością, wówczas nie można już ich odnieść do określonej osoby fizycznej. Sama randomizacja nie ograniczy szczególnego charakteru każdego zapisu, ponieważ każdy zapis nadal będzie pochodził od jednej osoby, której dane dotyczą, ale technika ta może chronić przed atakami / czynnikami ryzyka opartymi na wnioskowaniu i może być połączona z technikami uogólniania w celu zapewnienia silniejszych gwarancji prywatności. Wymagane mogą być dodatkowe techniki w celu zapewnienia, aby zapis nie umożliwiał zidentyfikowania jednej osoby fizycznej.

3.1.1. Dodawanie zakłóceń

Technika dodawania zakłóceń jest użyteczna zwłaszcza wtedy, gdy atrybuty mogą mieć istotny niekorzystny skutek dla poszczególnych osób fizycznych, i polega na modyfikowaniu atrybutów w zbiorze danych w taki sposób, aby były one mniej dokładne, przy jednoczesnym zachowaniu ogólnej dystrybucji. Podczas przetwarzania zbioru danych obserwator zakłada, że wartości są dokładne, ale będzie to prawdą tylko w pewnym stopniu. Przykładowo, jeżeli wzrost danej osoby fizycznej został pierwotnie zmierzony z dokładnością co do centymetra, zanonimizowany zbiór danych może zawierać informacje dotyczące wzrostu z dokładnością tylko do +/- 10 centymetrów. Jeżeli przedmiotową technikę stosuje się skutecznie, osoba trzecia nie będzie w stanie zidentyfikować żadnej osoby fizycznej, ani nie powinna być w stanie naprawić danych lub w inny sposób wykryć, w jaki sposób zmodyfikowano dane.

Dodawanie zakłóceń z reguły wymaga połączenia z innymi technikami anonimizacji, takimi jak usunięcie oczywistych atrybutów i *quasi*-identyfikatorów. Poziom zakłóceń powinien

zależeć od konieczności wymaganego poziomu informacji i wpływu ujawnienia chronionych atrybutów na prywatność poszczególnych osób fizycznych.

3.1.1.1. Gwarancje

- *Wyodrębnienie:* nadal możliwe jest wyodrębnienia zapisów konkretnej osoby fizycznej (być może w sposób uniemożliwiający identyfikację), chociaż zapisy są mniej wiarygodne.
- *Możliwość tworzenia powiązań:* nadal możliwe jest powiązanie zapisów tej samej osoby fizycznej, ale zapisy są mniej wiarygodne, przez co prawdziwy zapis może zostać powiązany z zapisem sztucznie dodanym (tj. „zakłócającym"). W niektórych przypadkach błędne przypisanie może narazić osobę, której dane dotyczą, na znaczący i nawet wyższy poziom ryzyka niż prawidłowe przypisanie.
- *Wnioskowanie:* mogą być możliwe ataki oparte na wnioskowaniu, ale wskaźnik powodzenia będzie niższy, a niektóre błędne akceptacje (i błędne odrzucenia) są wiarygodne.

3.1.1.2. Powszechne błędy

- Dodawanie niespójnych zakłóceń: jeżeli zakłócenie jest semantycznie niewykonalne (tj. wykracza „poza zakres" i nie przestrzega logiki między atrybutami w zbiorze), atakujący, mając dostęp do bazy danych, będzie w stanie odfiltrować zakłócenie i w niektórych przypadkach odtworzyć brakujące wejścia. Co więcej, jeżeli zbiór danych jest zbyt rozproszony[12], nadal może istnieć możliwość powiązania zakłóconych wpisów danych ze źródłem zewnętrznym.
- Założenie, że dodanie zakłóceń jest wystarczające: dodawanie zakłóceń jest środkiem uzupełniającym, który utrudnia atakującemu uzyskanie danych osobowych. Nie należy zakładać, że dodawanie zakłóceń stanowi samodzielne rozwiązanie umożliwiające anonimizację, chyba że ilość zakłóceń przewyższa ilości informacji zawartych w zbiorze danych.

3.1.1.3. Niepowodzenia w dodawaniu zakłóceń

Bardzo znanym eksperymentem w zakresie ponownej identyfikacji jest eksperyment przeprowadzony na bazie danych z danymi klientów dostawcy treści wideo – przedsiębiorstwa Netflix. Badacze przeprowadzili analizę właściwości geometrycznych tej bazy danych obejmującej ponad 100 mln ocen w skali 1–5 wyrażonych przez prawie 500 000 użytkowników w odniesieniu do ponad 18 000 filmów; baza ta została udostępniona publicznie przez przedsiębiorstwo po „zanonimizowaniu" jej zgodnie z wewnętrzną polityką prywatności i usunięciu wszystkich informacji umożliwiających identyfikację klienta poza ocenami i datami. Dodano zakłócenia polegające na nieznacznym podniesieniu lub obniżeniu ocen.

Mimo tego okazało się, że w zbiorze danych można jednoznacznie zidentyfikować 99 % zapisów użytkowników, wykorzystując jako kryteria wyboru 8 ocen i dat z błędami w zakresie 14 dni, natomiast obniżenie kryteriów wyboru (2 oceny i błąd w zakresie 3 dni) nadal umożliwia zidentyfikowanie 68 % użytkowników[13].

[12] Pojęcie to omówiono dalej w załączniku, s. 30.
[13] Narayanan, A., & Shmatikov, V. (2008 r., maj). Robust de-anonymization of large sparse datasets.[w:] *Security and Privacy, 2008 r. SP 2008. IEEE Symposium on* (s. 111–125). IEEE.

3.1.2. Permutacja

Technika ta, polegająca na tasowaniu wartości atrybutów w tabeli, tak aby niektóre z nich były sztucznie powiązane z różnymi osobami, których dane dotyczą, jest użyteczna, w przypadku gdy istotne jest zachowanie dokładnej dystrybucji każdego atrybutu w zbiorze danych.

Permutację można uznać za szczególną formę dodawania zakłóceń. W klasycznej technice dodawania zakłóceń atrybuty są modyfikowane za pomocą wartości randomizowanych. Generowanie spójnych zakłóceń może być trudnym zadaniem, a nieznaczne modyfikowanie wartości atrybutów może nie zapewniać odpowiedniej prywatności. Przy zastosowaniu technik permutacji jako rozwiązania alternatywnego zmienia się wartości w zbiorze danych poprzez podstawianie wartości z jednego zapisu do innego zapisu. Takie podstawianie zapewni, aby zakres i dystrybucja wartości pozostały takie same, ale korelacje między wartościami i poszczególnymi osobami fizycznymi były inne. Jeżeli między co najmniej dwoma atrybutami występuje związek logiczny lub korelacja statystyczna i atrybuty te zostały niezależnie permutowane, związek taki zostanie zlikwidowany. Przeprowadzenie permutacji zbioru powiązanych atrybutów może zatem być istotne, by nie naruszyć związku logicznego; w innym przypadku atakujący mógłby zidentyfikować permutowane atrybuty i odwrócić permutację.

Przykładowo, jeżeli weźmie się pod uwagę podzbiór atrybutów w zbiorze danych medycznych takich jak „przyczyny hospitalizacji/objawy/oddział odpowiadający", w większości przypadków wartości będą powiązane silnym związkiem logicznym, przez co wykryta zostałaby permutacja tylko jednej z tych wartości i istniałaby nawet możliwość jej odwrócenia.

Podobnie jak w przypadku dodawania zakłóceń sama permutacja nie zapewni anonimizacji i zawsze powinna być połączona z usuwaniem oczywistych atrybutów/*quasi*-identyfikatorów.

3.1.2.1. Gwarancje

- *Wyodrębnienie:* podobnie jak w przypadku dodawania zakłóceń nadal możliwe jest wyodrębnienie zapisów konkretnej osoby fizycznej, ale zapisy są mniej wiarygodne.
- *Możliwość tworzenia powiązań:* jeżeli permutacja wpływa na atrybuty i *quasi*-identyfikatory, może ona uniemożliwić „prawidłowe" powiązanie atrybutów ze zbiorem danych zarówno wewnętrznie, jak i zewnętrzne, ale nadal pozwala ona na „nieprawidłową" możliwość tworzenia powiązań, ponieważ prawdziwy wpis można powiązać z różnymi osobami, których dane dotyczą.
- *Wnioskowanie:* nadal można wyciągać wnioski ze zbioru danych, w szczególności jeżeli atrybuty są skorelowane lub występuje między nimi silny związek logiczny; nie wiedząc, które atrybuty objęto permutacją, atakujący musi jednak uznać, że jego wnioskowanie opiera się na złej hipotezie, w związku z czym możliwe jest tylko wnioskowanie probabilistyczne.

3.1.2.2. Powszechne błędy

- Wybieranie niewłaściwego atrybutu: permutowanie niechronionych szczególnie i nieobciążonych ryzykiem atrybutów nie skutkowałoby znaczącym zyskiem pod względem ochrony danych osobowych. W praktyce, jeżeli szczególnie chronione/obciążone ryzykiem atrybuty nadal byłyby związane z atrybutami

pierwotnymi, atakujący wciąż byłby w stanie wyciągnąć dane szczególnie chronione dotyczące poszczególnych osób fizycznych.

- Losowe permutowanie atrybutów: jeżeli dwa atrybuty są ściśle skorelowane, losowe permutowanie atrybutów nie zapewni silnych gwarancji. Ten powszechny błąd zilustrowano w tabeli 1.
- Zakładanie, że permutacja wystarcza: podobnie jak w przypadku dodawania zakłóceń sama permutacja nie zapewnia anonimowości, dlatego należy ją łączyć z innymi technikami takimi jak usuwanie oczywistych atrybutów.

3.1.2.3. Niepowodzenia w zakresie permutacji

Przykład ten pokazuje, w jaki sposób losowe permutowanie atrybutów skutkuje słabymi gwarancjami prywatności, w przypadku gdy między różnymi atrybutami istnieją powiązania logiczne. W następstwie podjętej próby anonimizacji bardzo łatwo jest wydedukować dochód każdej osoby fizycznej w zależności od jej stanowiska pracy (i roku urodzenia). Na przykład na podstawie bezpośredniej kontroli danych można stwierdzić, że dyrektor generalny uwzględniony w tabeli prawdopodobnie urodził się w 1957 r. i otrzymuje najwyższe wynagrodzenie, natomiast bezrobotny urodził się w 1964 r., a jego dochód jest najniższy.

Rok:	Płeć	Stanowisko pracy	Dochód (permutowany)
1957	M	Inżynier	70 tys.
1957	M	Dyrektor generalny	5 tys.
1957	M	Bezrobotny	43 tys.
1964	M	Inżynier	100 tys.
1964	M	Manager	45 tys.

Tabela 1. Przykład nieskutecznej anonimizacji przez permutację skorelowanych atrybutów.

3.1.3. Prywatność różnicowa

Prywatność różnicowa[14] należy do rodziny technik randomizacji, ale opiera się na innym podejściu: o ile w praktyce dodawanie zakłóceń odbywa się, zanim zbiór danych ma zostać udostępniony, o tyle prywatność różnicową można zastosować w czasie, gdy administrator danych generuje zanonimizowane widoki zbioru danych, jednocześnie zachowując kopie danych pierwotnych. Takie zanonimizowane widoki zwykle generuje się przez podzbiór zapytań na potrzeby określonej osoby trzeciej. Podzbiór ten obejmuje pewne losowe zakłócenia dodane celowo po przeprowadzeniu anonimizacji. Dzięki prywatności różnicowej administrator danych wie, jak wiele zakłóceń musi dodać i w jakiej formie, aby uzyskać niezbędne gwarancje prywatności[15]. W tym kontekście niezwykle ważne będzie ciągłe monitorowanie (przynajmniej w odniesieniu do każdego nowego zapytania) pod kątem jakiejkolwiek możliwości zidentyfikowania osoby fizycznej w zbiorze rezultatów zapytania. Należy jednak wyjaśnić, że techniki prywatności różnicowej nie zmienią danych pierwotnych i dlatego, dopóki istnieją dane pierwotne, administrator danych jest w stanie zidentyfikować

[14] Dwork, C. (2006). Differential privacy. [w:] *Automata, languages and programming* (s. 1–12). Springer Berlin Heidelberg.
[15] Por. Ed Felten (2012) Protecting privacy by adding noise. Adres URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/

poszczególne osoby fizyczne na skutek zapytań w ramach prywatności różnicowej, biorąc pod uwagę wszystkie sposoby, jakimi można się posłużyć. Takie rezultaty należy uznać za dane osobowe.

Jedną z korzyści podejścia opartego na prywatności różnicowej jest fakt, że zbiory danych udostępniane są upoważnionym osobom trzecim w odpowiedzi na szczególne zapytanie, a nie przez udostępnianie jednego zbioru danych. Aby pomóc w kontroli, administrator danych może zachować listę wszystkich zapytań i próśb, zapewniając, aby osoby trzecie nie miały dostępu do danych, do których nie są upoważnione. Zapytanie może również zostać objęte technikami anonimizacji, w tym dodawaniem zakłóceń lub zastąpieniem w celu dalszej ochrony prywatności. Nadal otwartą kwestią badawczą pozostaje znalezienie dobrego, interaktywnego mechanizmu zapytanie-odpowiedź, który potrafi też odpowiadać na wszelkie pytania dość dokładnie (czyli w sposób mniej zakłócony), jednocześnie zachowując prywatność.

Aby ograniczyć ataki oparte na wnioskowaniu i możliwości tworzenia powiązań, konieczne jest prowadzenie monitorowania zapytań wprowadzanych przez podmiot i obserwowanie zdobytych informacji na temat osób, których dane dotyczą; dlatego też nie należy udostępniać baz danych opartych na „prywatności różnicowej" w otwartych wyszukiwarkach internetowych, które nie oferują identyfikowalności podmiotów wprowadzających zapytania.

3.1.3.1. Gwarancje

- *Wyodrębnienie:* gdy wynikiem są tylko statystyki, a przepisy mające zastosowanie do zbioru zostały dobrze wybrane, wykorzystanie odpowiedzi do wyodrębnienia konkretnej osoby fizycznej powinno być niemożliwe.
- *Możliwość tworzenia powiązań:* możliwe jest powiązanie wpisów odnoszących się do określonej osoby fizycznej między dwiema odpowiedziami poprzez wykorzystanie wielu zapytań.
- *Wnioskowanie:* istnieje możliwość wywnioskowania informacji na temat poszczególnych osób fizycznych lub grup poprzez wykorzystanie wielu zapytań.

3.1.3.2. Powszechne błędy

- Dodanie niewystarczającego zakłócenia: aby uniemożliwić powiązanie z podstawową wiedzą, wyzwaniem jest dostarczenie minimalnej liczby dowodów na to, czy określona osoba lub grupa osób, których dane dotyczą, wniosły wkład do tego zbioru danych. Największą trudnością pod względem ochrony danych jest zdolność do wygenerowania odpowiedniej ilości zakłóceń dodawanych do prawdziwych odpowiedzi, tak aby chronić prywatność osób fizycznych, przy jednoczesnym zachowaniu użyteczności udostępnionych odpowiedzi.

3.1.3.3 Niepowodzenia w zakresie prywatności różnicowej

Traktowanie każdego zapytania oddzielnie: połączenie rezultatów zapytań może umożliwić ujawnienie informacji, które z założenia miały być tajne. Jeżeli nie zachowano historii zapytań, atakujący może konstruować wiele pytań do bazy danych opartej na „prywatności różnicowej", które stopniowo ograniczają amplitudę wynikowych prób do momentu, w którym może wyłonić się w sposób deterministyczny lub z dużym prawdopodobieństwem szczególny charakter jednej osoby lub grupy osób, których dane dotyczą. Ponadto dodatkowym zastrzeżeniem jest unikanie błędu, jakim jest zakładanie,

że dane są anonimowe dla osoby trzeciej, jeżeli administrator danych nadal może zidentyfikować osobę, której dane dotyczą, w pierwotnej bazie danych, biorąc pod uwagę wszystkie sposoby, jakimi można się posłużyć.

3.2. Uogólnianie

Uogólnianie stanowi drugą rodzinę technik anonimizacji. Podejście to polega na uogólnianiu lub osłabianiu atrybutów osób, których dane dotyczą, poprzez modyfikowanie odpowiedniego zakresu lub rzędu wielkości (tj. raczej region, a nie miasto, raczej miesiąc, a nie tydzień). Chociaż uogólnianie może być skuteczne w uniemożliwianiu wyodrębniania, nie pozwala ono na skuteczną anonimizację we wszystkich przypadkach; w szczególności uogólnianie wymaga określonych i zaawansowanych podejść ilościowych w celu zapobieżenia możliwości tworzenia powiązań i wnioskowaniu.

3.2.1. Agregacja i k-anonimizacja

Techniki agregacji i k-anonimizacji mają na celu uniemożliwienie wyodrębnienia osoby, której dane dotyczą, poprzez zgrupowanie tych osób z co najmniej k innymi osobami fizycznymi. Aby to osiągnąć, uogólnia się wartości atrybutów do takiego zakresu, w jakim każdej osobie fizycznej przypisana jest ta sama wartość. Przykładowo przez obniżenie poziomu szczegółowości lokalizacji z miasta do państwa baza danych obejmuje większą liczbę osób, których dane dotyczą. Poszczególne daty urodzenia mogą zostać uogólnione do przedziału dat lub pogrupowane według miesięcy lub lat. Inne atrybuty numeryczne (np. wynagrodzenie, waga, wzrost lub dawka leku) mogą zostać uogólnione przez zastosowanie wartości przedziałowych (np. wynagrodzenie 20 000–30 000 EUR). Metody te mogą być stosowane, w przypadkach w których korelacja wartości punktowych atrybutów może utworzyć *quasi*-identyfikatory.

3.2.1.1. Gwarancje

- *Wyodrębnienie:* ponieważ te same atrybuty są teraz wspólne dla k użytkowników, wyodrębnienie konkretnej osoby fizycznej z grupy k użytkowników powinno być już niemożliwe.
- *Możliwość tworzenia powiązań:* chociaż możliwość tworzenia powiązań jest ograniczona, nadal możliwe jest powiązanie zapisów według grup k użytkowników. Wtedy w ramach danej grupy prawdopodobieństwo tego, że dwa zapisy odpowiadają temu samemu pseudoidentyfikatorowi wynosi 1/k (i może być znacznie wyższe niż prawdopodobieństwo, że takich wpisów nie można powiązać).
- *Wnioskowanie:* główną wadą modelu k-anonimizacji jest to, że nie zapobiega on żadnemu rodzajowi ataków opartych na wnioskowaniu. W praktyce, jeżeli k osób fizycznych należy do tej samej grupy, wówczas jeżeli wiadomo, do której grupy należy dana osoba, bardzo łatwo uzyskać wartość tej właściwości.

3.2.1.2. Powszechne błędy

- Brak niektórych *quasi*-identyfikatorów: istotnym parametrem przy rozważaniu k-anonimizacji jest próg k. Im większa wartość k, tym silniejsze są gwarancje prywatności. Powszechnym błędem jest sztuczne zwiększanie wartości k przez ograniczanie uwzględnianego zbioru *quasi*-identyfikatorów. Ograniczanie *quasi*-identyfikatorów ułatwia budowanie klastrów k-użytkowników dzięki właściwej mocy identyfikowania przypisanej innym atrybutom (w szczególności, jeżeli niektóre z nich

są szczególnie chronione lub charakteryzuje je bardzo wysoka entropia, jak w przypadku bardzo rzadkich atrybutów). Nieuwzględnienie wszystkich *quasi*-identyfikatorów przy wyborze atrybutu do uogólnienia jest poważnym błędem; jeżeli możliwe jest wykorzystanie określonych atrybutów w celu wyodrębnienia konkretnej osoby fizycznej w klastrze k użytkowników, wówczas uogólnienie nie chroni niektórych osób (zob. przykład w tabeli 2).

- Mała wartość k: podobnie problem stanowi dążenie do uzyskania małej wartości k. Jeżeli k jest zbyt małe, waga każdej osoby fizycznej w klastrze jest zbyt znacząca, a ataki oparte na wnioskowaniu odnotowują wyższy wskaźnik skuteczności. Na przykład, jeżeli k=2, wówczas prawdopodobieństwo, że dwie osoby dzielą tę samą właściwość jest większe niż w przypadku k>10.
- Niegrupowanie osób o tej samej wadze: grupowanie zbioru osób fizycznych o nierównej dystrybucji atrybutów również może stanowić problem. Wpływ zapisu danej osoby fizycznej na zbiór danych będzie się różnił: niektóre zapisy będą stanowiły znaczący odsetek wpisów, natomiast wkłady innych zapisów pozostaną raczej nieznaczące. Dlatego też ważne jest upewnienie się, że k jest wystarczająco wysokie, tak aby żadna osoba fizyczna nie reprezentowała zbyt dużego odsetka wpisów w klastrze.

3.1.3.3. Niepowodzenia w zakresie k-anonimizacji

Głównym problem związanym z k-anonimizacją jest to, że nie zapobiega ona atakom opartym na wnioskowaniu. W podanym przykładzie, jeżeli atakujący wie, że konkretna osoba fizyczna jest objęta zbiorem danych i że urodziła się w 1964 r., wie także, że osoba ta miała zawał serca. Ponadto, jeżeli wiadomo, że ten zbiór danych otrzymano od francuskiej organizacji, oznacza to, że każda osoba fizyczna mieszka w Paryżu, ponieważ pierwsze trzy cyfry paryskiego kodu pocztowego to 750*.

Rok:	Płeć	Kod pocztowy	Diagnoza
1957	M	750*	Zawał serca
1957	M	750*	Cholesterol
1957	M	750*	Cholesterol
1964	M	750*	Zawał serca
1964	M	750*	Zawał serca

Tabela 2. Przykład słabo skonstruowanej k-anonimizacji.

3.2.2. L-dywersyfikacja/t-bliskość

L-dywersyfikacja stanowi rozszerzenie k-anonimizacji w celu zapewnienia, aby deterministyczne ataki oparte na wnioskowaniu nie były już możliwe poprzez zagwarantowanie, że w każdej klasie równoważności każdy atrybut ma co najmniej *l* różnych wartości.

Jednym podstawowym celem do osiągnięcia jest ograniczenie występowania klas równoważności o słabej zmienności atrybutów, tak aby atakujący posiadający podstawową wiedzę na temat określonej osoby, której dane dotyczą, zawsze miał duży stopień niepewności.

L-dywersyfikacja jest użyteczna do celów ochrony danych przed atakami opartymi na wnioskowaniu, gdy wartości atrybutów są dobrze rozmieszczone. Należy jednak podkreślić, że technika ta nie może zapobiec wyciekowi informacji, jeżeli atrybuty są nierówno rozmieszczone w ramach przedziału bądź należą do małego zakresu wartości lub znaczeń semantycznych. Ostatecznie l-dywersyfikacja podlega atakom opartym na wnioskowaniu probabilistycznym.

T-bliskość stanowi udoskonalenie l-dywersyfikacji pod tym względem, że ma na celu utworzenie równoważnych klas, które odzwierciedlają początkową dystrybucję atrybutów w tabeli. Technika ta jest użyteczna, gdy istotne jest zachowanie danych możliwie najbliżej danych pierwotnych; w tym celu nakłada się dalsze ograniczenie na klasę równoważności, mianowicie, że nie tylko powinno istnieć co najmniej *l* różnych wartości w ramach każdej klasy równoważności, ale również, że każda wartość jest reprezentowana tyle razy, ile jest konieczne, aby odzwierciedlić początkową dystrybucję każdego atrybutu.

3.2.2.1. Gwarancje

- *Wyodrębnienie:* podobnie jak w przypadku k-anonimizacji l-dywersyfikacja i t-bliskość mogą zapewnić, aby zapisy dotyczące danej osoby fizycznej nie mogły zostać wyodrębnione w bazie danych.
- *Możliwość tworzenia powiązań:* l-dywersyfikacja i t-bliskość nie stanowią udoskonalenia w porównaniu z k-anonimizacją, jeżeli chodzi o możliwość tworzenia powiązań. Kwestia ta prezentuje się tak samo jak w przypadku każdego innego klastra: prawdopodobieństwo, że te same wpisy należą do tej samej osoby, której dane dotyczą, jest wyższe niż 1/N (gdzie N oznacza liczbę osób, których dane dotyczą, występujących w bazie danych).
- *Wnioskowanie:* głównym udoskonaleniem l-dywersyfikacji i t-bliskości w porównaniu z k-anonimizacją jest to, że nie ma już możliwości organizowania ze 100 % pewnością ataków opartych na wnioskowaniu na bazę danych objętą l-dywersyfikacją i t-bliskością.

3.2.2.2. Powszechne błędy

- Zabezpieczenie szczególnie chronionych wartości atrybutów poprzez zmieszanie ich z innymi atrybutami szczególnie chronionymi: dwie wartości atrybutu w klastrze nie wystarczają do zapewnienia gwarancji prywatności. W praktyce dystrybucja wartości szczególnie chronionych w każdym klastrze powinna przypominać dystrybucję tych wartości w całej populacji lub przynajmniej powinna być ona jednakowa w całym klastrze.

3.2.2.3. Niepowodzenia w zakresie l-dywersyfikacji

W poniższej tabeli przeprowadzono l-dywersyfikację w odniesieniu do atrybutu „diagnoza”; wiedząc, że osoba urodzona w 1964 r. została uwzględniona w tej tabeli, nadal można jednak stwierdzić z bardzo dużym prawdopodobieństwem, że miała ona zawał serca.

Rok:	Płeć	Kod pocztowy	Diagnoza
1957	M	750*	**Zawał serca**
1957	M	750*	**Cholesterol**
1957	M	750*	**Cholesterol**
1957	M	750*	**Cholesterol**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Cholesterol**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**
1964	M	750*	**Zawał serca**

Tabela 3. Tabela poddana l-dywersyfikacji, w której wartości „diagnozy” nie zostały równomiernie rozłożone.

Nazwisko	Data urodzenia	Płeć
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tabela 4. Wiedząc, że osoby te zostały uwzględnione w tabeli 3, atakujący mógł wywnioskować, że miały one zawał serca.

4. Pseudonimizacja

Pseudonimizacja polega na zastępowaniu jednego atrybutu (z reguły atrybutu nietypowego) w zapisie innym atrybutem. W związku z tym nadal istnieje prawdopodobieństwo pośredniego zidentyfikowania osoby fizycznej; dlatego też stosowanie samej pseudonimizacji nie będzie skutkowało anonimowym zbiorem danych. W niniejszej opinii omówiono jednak przedmiotową technikę ze względu na wiele związanych z nią mylnych przekonań i błędów.

Pseudonimizacja ogranicza możliwość tworzenia powiązań zbioru danych z prawdziwą tożsamością osoby, której dane dotyczą; technika ta stanowi zatem użyteczny środek bezpieczeństwa, ale nie metodę anonimizacji.

Wynik pseudonimizacji może być niezależny od wartości początkowej (jak w przypadku numerów losowych generowanych przez administratora danych lub nazwiska wybranego przez osobę, której dane dotyczą) lub może opierać się na pierwotnych wartościach atrybutu lub zbioru atrybutów, np. funkcji skrótu lub układu szyfrowania.

Do najczęściej stosowanych technik pseudonimizacji należą:

- szyfrowanie z kluczem tajnym: w tym przypadku posiadacz klucza może z łatwością ponownie zidentyfikować każdą osobę, której dane dotyczą, poprzez odszyfrowanie zbioru danych, ponieważ dane osobowe nadal znajdują się w tym zbiorze danych, chociaż w zaszyfrowanej formie. Zakładając, że zastosowano układ szyfrowania w oparciu o stan wiedzy naukowej i technicznej, możliwość odszyfrowania istnieje wyłącznie w przypadku, gdy znany jest klucz;
- funkcja skrótu: oznacza funkcję, która z wkładu każdej wielkości (wkład może być jednym atrybutem lub zbiorem atrybutów) daje wynik stałej wielkości i której nie można odwrócić; oznacza to, że ryzyko odwrócenia, występujące przy szyfrowaniu, już nie istnieje. Jeżeli znany jest jednak zakres wartości wkładu przy funkcji skrótu, będzie on mógł zostać ponownie odtworzony za pomocą funkcji skrótu w celu uzyskania prawidłowej wartości dla konkretnego zapisu. Na przykład, jeżeli zbiór danych został poddany pseudonimizacji poprzez skrócenie krajowego numeru identyfikacyjnego, można uzyskać ten numer, po prostu skracając wszystkie możliwe wartości wkładu i porównując wyniki z odpowiednimi wartościami ze zbioru danych. Funkcje skrótu są zwykle opracowane w taki sposób, aby można było prowadzić stosunkowo szybkie obliczenia i są przedmiotem ataków siłowych[16]. Można również utworzyć wstępnie obliczone tabele, aby umożliwić odwrócenie masowe dużego zbioru wartości skrótu.

  Stosowanie funkcji skrótu z losowym ciągiem znaków (w której do skracanego atrybutu dodaje się losowy ciąg znaków, ang. „salt") może ograniczyć prawdopodobieństwo uzyskania wartości wkładu, ale obliczanie wartości pierwotnego atrybutu ukrytej za wynikiem funkcji skrótu z losowym ciągiem może jednak nadal być wykonalne w ramach uzasadnionych środków[17];

[16] Ataki takie polegają na wypróbowaniu wszystkich możliwych kombinacji w celu utworzenia tabel korelacji.
[17] W szczególności, jeżeli znany jest rodzaj atrybutu (nazwisko, numer ubezpieczenia społecznego, data urodzenia itp.). W celu dodania wymogu obliczeniowego można opierać się na funkcji skrótu z wyprowadzaniem klucza, w której wyliczona wartość zostaje kilkukrotnie skrócona za pomocą krótkiego losowego ciągu znaków.

- funkcja skrótu z kluczem, w przypadku której klucz jest przechowywany: oznacza określoną funkcję skrótu, która wykorzystuje klucz tajny jako dodatkowy wkład (różni się ona od funkcji skrótu z losowym ciągiem znaków, ponieważ zazwyczaj ciąg losowy nie stanowi tajemnicy). Administrator danych może ponownie odtworzyć tę funkcję względem atrybutu, wykorzystując klucz tajny, ale znacznie trudniej jest odtworzyć ją atakującemu bez znajomości klucza, ponieważ liczba możliwości, które trzeba sprawdzić, jest na tyle duża, że ich sprawdzanie jest niepraktyczne;
- szyfrowanie deterministyczne lub funkcja skrótu z kluczem, w przypadku której klucz jest usuwany: technika ta może być utożsamiona z wybieraniem losowego numeru jako pseudonimu dla każdego atrybutu w bazie danych, a następnie z usuwaniem tabeli korelacji. Rozwiązanie to pozwala[18] ograniczyć ryzyko możliwości tworzenia powiązań między danymi osobowymi w zbiorze danych a danymi odnoszącymi się do tej samej osoby fizycznej w innym zbiorze danych, w którym używa się innego pseudonimu. Rozważając algorytm oparty na stanie wiedzy naukowej i technicznej, odszyfrowanie lub odtworzenie funkcji będzie dla atakującego trudne pod względem obliczeniowym, ponieważ wiązałoby się ono ze sprawdzaniem każdego możliwego klucza ze względu na fakt, iż klucz ten nie jest dostępny;
- tokenizacja: technika ta jest zwykle stosowana w sektorze finansowym (chociaż nie tylko) w celu zastąpienia numerów identyfikacyjnych kart wartościami, które ograniczają użyteczność dla atakującego. Opiera się ona na omówionych wcześniej technikach i zwykle polega na stosowaniu mechanizmów szyfrowania jednokierunkowego lub na przypisaniu, za pomocą funkcji indeksu, sekwencji liczb lub losowo wygenerowanych liczb, które nie zostały w sposób matematyczny uzyskane z danych pierwotnych.

4.1. Gwarancje

- *Wyodrębnienie:* nadal możliwe jest wyodrębnienie zapisów poszczególnych osób, ponieważ dana osoba wciąż może zostać zidentyfikowana przez atrybut nietypowy, który wynika z funkcji pseudonimizacji (= atrybut pseudonimiczny).
- *Możliwość tworzenia powiązań:* nadal łatwo będzie można tworzyć powiązania między zapisami, wykorzystując ten sam atrybut pseudonimiczny w celu odniesienia się do tej samej osoby fizycznej. Mimo że do tej samej osoby fizycznej, której dane dotyczą, wykorzystuje się różne atrybuty pseudonimiczne, nadal będzie istniała możliwość tworzenia powiązań za pomocą innych atrybutów. Jedynie w przypadku, gdy nie można wykorzystać żadnego innego atrybutu w zbiorze danych w celu zidentyfikowania osoby, której dane dotyczą, i jeżeli każde powiązanie między atrybutem pierwotnym a atrybutem pseudonimicznym zostało zlikwidowane (w tym przez usunięcie danych pierwotnych), nie będzie żadnych oczywistych odniesień między dwoma zbiorami danych wykorzystującymi różne atrybuty pseudonimiczne.
- *Wnioskowanie:* ataki oparte na wnioskowaniu ukierunkowane na prawdziwą tożsamość osoby, której dane dotyczą, są możliwie w ramach zbioru danych lub między różnymi zbiorami danych, które wykorzystują te same atrybuty pseudonimiczne w odniesieniu do określonej osoby fizycznej, lub jeżeli pseudonimy nie wymagają wyjaśnienia i nie ukrywają w odpowiedni sposób prawdziwej tożsamości osoby, której dane dotyczą.

[18] W zależności od innych atrybutów w zbiorze danych i od usunięcia danych pierwotnych.

4.2. Powszechne błędy

- Przekonanie, że dane opatrzone pseudonimem zostały zanonimizowane: administratorzy danych często zakładają, że usunięcie lub zastąpienie jednego atrybutu lub ich większej liczby wystarcza do zanonimizowania zbioru danych. Wiele przykładów pokazało, że tak nie jest; zwykła zmiana identyfikatora danych osobowych nie uniemożliwia zidentyfikowania osoby, której dane dotyczą, jeżeli w zbiorze danych pozostają *quasi*-identyfikatory lub jeżeli wartości innych atrybutów nadal umożliwiają zidentyfikowanie danej osoby fizycznej. W wielu przypadkach zidentyfikowanie konkretnej osoby fizycznej w zbiorze danych opatrzonych pseudonimem jest tak proste, jak w przypadku danych pierwotnych. Należy podjąć dodatkowe działania w celu uznania zbioru danych za zanonimizowany, w tym usunąć i uogólnić atrybuty lub wykasować dane pierwotne lub przynajmniej sprawić, by były one w dużym stopniu zagregowane.
- Powszechne błędy przy stosowaniu pseudonimizacji jako techniki ograniczania możliwości tworzenia powiązań:
    - wykorzystywanie tego samego klucza w różnych bazach danych: eliminowanie możliwości tworzenia powiązań różnych zbiorów danych polega w dużym stopniu na stosowaniu algorytmu kluczy oraz na fakcie, że jedna osoba będzie odpowiadała różnym atrybutom pseudonimicznym w różnych kontekstach. Dlatego, aby ograniczyć możliwości tworzenia powiązań, ważne jest unikanie wykorzystywania tego samego klucza w różnych bazach danych;
    - wykorzystywanie różnych kluczy („kluczy zmienianych") w odniesieniu do różnych użytkowników: może powstać pokusa wykorzystywania różnych kluczy w odniesieniu do różnych zbiorów użytkowników i zmiany klucza przy każdym użyciu (na przykład, użycie tego samego klucza do zarejestrowania 10 wpisów dotyczących tego samego użytkownika). Jeżeli takie działanie nie jest odpowiednio zaprojektowane, może jednak spowodować wystąpienie wzorców, co częściowo ograniczy zamierzone korzyści. Przykładowo zmienianie klucza w drodze stosowania szczególnych zasad w odniesieniu do konkretnych osób fizycznych ułatwiłoby możliwość tworzenia powiązań wpisów odpowiadających danym osobom. Również znikanie powtarzających się danych opatrzonych pseudonimem z bazy danych w czasie, gdy pojawiają się nowe, może sygnalizować, że oba zapisy odnoszą się do tej samej osoby fizycznej;
    - zachowywanie klucza: jeżeli klucz tajny jest przechowywany razem z danymi opatrzonymi pseudonimem, a dane zostają narażone na szwank, atakujący może być w stanie z łatwością powiązać dane pseudonimiczne z ich pierwotnymi atrybutami. Ta sama zasada ma zastosowanie, gdy klucz jest przechowywany oddzielnie od danych, ale nie w sposób bezpieczny.

4.3. Wady pseudonimizacji

- *Opieka zdrowotna*

1. Nazwisko, adres, data urodzenia	2. Okres wypłacania świadczenia z tytułu pomocy specjalnej	3. Współczynnik masy ciała	6. Nr referencyjny badanej kohorty
[redacted]	< 2 lata	15	QA5FRD4
	> 5 lat	14	2B48HFG
	< 2 lata	16	RC3URPQ
	> 5 lat	18	SD289K9
	< 2 lata	20	5E1FL7Q

Tabela 5. Przykład pseudonimizacji przez skracanie (nazwisko, adres, data urodzenia), którą można łatwo odwrócić.

Zbiór danych został utworzony w celu zbadania związku między wagą danej osoby a otrzymywaniem płatności świadczenia z tytułu pomocy specjalnej. Pierwotny zbiór danych obejmował nazwisko, adres i datę urodzenia osób, których dane dotyczą, ale zostały one usunięte. Numer referencyjny badanej kohorty został przyznany z usuniętych danych za pomocą funkcji skrótu. Chociaż nazwisko, adres i data urodzenia zostały usunięte z tabeli, to jeżeli znane są nazwisko, adres i data urodzenia osoby, której dane dotyczą, oraz znana jest zastosowana funkcja skrótu, obliczenie numerów referencyjnych badanej kohorty jest bardzo łatwe.

- *Sieci społecznościowe*

Wykazano[19], że szczególnie chronione informacje na temat określonych osób fizycznych można wydobyć z grafów powiązań społecznościowych, mimo technik „pseudonimizacji" zastosowanych do takich danych. Dostawca sieci społecznościowej błędnie zakładał, że pseudonimizacja była wystarczająca, aby zapobiec identyfikacji po sprzedaży danych innym przedsiębiorstwom do celów marketingowych i reklamowych. Zamiast prawdziwych nazwisk dostawca użył pseudonimów, ale to zdecydowanie nie wystarczyło do anonimizacji wykorzystanych profili, ponieważ związki między różnymi osobami fizycznymi są niepowtarzalne i mogą zostać wykorzystane jako identyfikatory.

- *Lokalizacje*

Badacze z Instytutu Technologii w Massachusetts (MIT)[20] przeprowadzili ostatnio analizę zbioru danych opatrzonych pseudonimem składającego się z danych z 15 miesięcy przedstawiających współrzędne dotyczące czasowej mobilności przestrzennej 1,5 mln ludzi na obszarze o promieniu 100 km. Badacze wykazali, że za pomocą czterech punktów lokalizacji można wyodrębnić 95 % populacji i że zaledwie dwa punkty były wystarczające, aby wyodrębnić ponad 50 % osób, których dane dotyczą (przy czym jeden z tych punktów jest znany i jest to najprawdopodobniej „dom" lub „biuro"), w bardzo dużym stopniu ograniczając ochronę prywatności, chociaż tożsamości poszczególnych osób fizycznych zostały opatrzone pseudonimem przez zastąpienie ich prawdziwych atrybutów […] innymi oznaczeniami.

[19] A. Narayanan and V. Shmatikov, „De-anonymizing social networks", [w:] 30th IEEE Symposium on Security and Privacy, 2009 r.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen i V. Blondel, „Unique in the Crowd: The privacy bounds of human mobility", Nature, nr 1376, 2013 r.

5. Wnioski i zalecenia

5.1. Wnioski

Techniki anonimizacji są przedmiotem intensywnych badań, a w niniejszej opinii spójnie wykazano, że każda technika ma swoje zalety i wady. W większości przypadków nie jest możliwe udzielenie minimalnych zaleceń w odniesieniu do parametrów, które należy zastosować, ponieważ każdy zbiór danych należy rozpatrywać indywidualnie.

W wielu przypadkach zanonimizowany zbiór danych nadal może stanowić ryzyko szczątkowe dla osób, których dane dotyczą. W praktyce, chociaż nie jest już możliwe dokładne odzyskanie zapisu danej osoby fizycznej, nadal może być możliwe zdobycie informacji na temat tej osoby za pomocą innych dostępnych (publicznie lub niepublicznie) źródeł informacji. Należy podkreślić, że poza bezpośrednim wpływem skutków niedostatecznego procesu anonimizacji na osoby, których dane dotyczą (rozdrażnienie, strata czasu, poczucie utraty kontroli przez bycie włączonym do klastra bez świadomości lub uprzedniej zgody), mogą wystąpić inne pośrednie skutki uboczne niedostatecznej anonimizacji w każdym przypadku, gdy osoba, której dane dotyczą, w wyniku przetwarzania zanonimizowanych danych została przez pomyłkę uwzględniona przez dowolnego atakującego jako cel – w szczególności jeżeli atakujący ma złe zamiary. Dlatego też grupa robocza podkreśla, że techniki anonimizacji mogą zapewnić gwarancje prywatności, ale jedynie wtedy, gdy stosowanie tych technik zostało odpowiednio zaprojektowane, co oznacza, że aby osiągnąć docelowy poziom anonimizacji, niezbędne jest jasne określenie warunków wstępnych (kontekst) i celu lub celów procesu anonimizacji.

5.2. Zalecenia

- Niektóre techniki anonimizacji wiążą się nieodłącznie z konkretnymi ograniczeniami. Administratorzy danych muszą poważnie rozważyć te ograniczenia, zanim zastosują daną technikę w celu przeprowadzenia procesu anonimizacji. Muszą oni zwrócić uwagę na cele, jakie należy osiągnąć przez anonimizację – takie jak ochrona prywatności osób fizycznych przy publikowaniu zbioru danych lub dopuszczaniu wyszukania części informacji ze zbioru danych.
- Żadna z technik opisanych w niniejszym dokumencie nie spełnia w całości wszystkich kryteriów skutecznej anonimizacji (tj. braku możliwości wyodrębnienia określonej osoby fizycznej, braku możliwości tworzenia powiązań między zapisami dotyczącymi określonej osoby i braku możliwości wnioskowania w odniesieniu do określonej osoby). Ponieważ niektóre z tych zagrożeń można jednak wyeliminować w pełni lub częściowo za pomocą konkretnej techniki, konieczne jest ostrożne projektowanie przy opracowywaniu zastosowania danej techniki do określonej sytuacji i przy stosowaniu połączenia tych technik jako sposobu zwiększenia niezawodności wyniku.

W poniższej tabeli przedstawiono przegląd zalet i wad technik pod względem trzech podstawowych wymogów:

	Czy nadal istnieje ryzyko wyodrębnienia?	Czy nadal istnieje ryzyko możliwości tworzenia powiązań?	Czy nadal istnieje ryzyko wnioskowania?
Pseudonimizacja	Tak	Tak	Tak
Dodawanie zakłóceń	Tak	Być może nie	Być może nie
Zastąpienie	Tak	Tak	Być może nie
Agregacja lub k-anonimizacja	Nie	Tak	Tak
L-dywersyfikacja	Nie	Tak	Być może nie
Prywatność różnicowa	Być może nie	Być może nie	Być może nie
Skracanie/Tokenizacja	Tak	Tak	Być może nie

Tabela 6. Zalety i wady rozważanych technik.

- O optymalnym rozwiązaniu należy decydować w odniesieniu do poszczególnych przypadków. Rozwiązanie (tj. proces całkowitej anonimizacji) spełniające przedmiotowe trzy kryteria chroniłoby przed identyfikacją przeprowadzaną w najbardziej prawdopodobne i uzasadnione sposoby, jakimi może posłużyć się administrator danych lub jakakolwiek osoba trzecia.
- W każdym przypadku, w którym propozycja nie spełnia jednego z powyższych kryteriów, należy przeprowadzić dokładną ocenę ryzyka identyfikacji. Ocenę tę należy przedstawić właściwemu organowi, jeżeli prawo krajowe wymaga, aby organ ten ocenił lub zatwierdził proces anonimizacji.

W celu ograniczenia ryzyka identyfikacji należy uwzględnić następujące dobre praktyki:

<u>Dobre praktyki anonimizacji:</u>

Ogólnie:

- nie należy opierać się na podejściu „udostępnij i zapomnij”. Ze względu na ryzyko szczątkowe identyfikacji administratorzy danych powinni:
    - 1. identyfikować nowe rodzaje ryzyka i regularnie przeprowadzać ponowne oceny ryzyka szczątkowego;
    - 2. ocenić, czy kontrole w odniesieniu do zidentyfikowanego ryzyka są wystarczające i odpowiednio dostosowywane; ORAZ
    - 3. monitorować i kontrolować ryzyko;
- w ramach takiego ryzyka szczątkowego należy uwzględnić potencjał identyfikacyjny niezanonimizowanej części zbioru danych (jeżeli taka istnieje), w szczególności jeżeli została ona połączona z częścią zanonimizowaną, oraz możliwe korelacje między atrybutami (np. między danymi dotyczącymi lokalizacji geograficznej a danymi dotyczącymi poziomu zamożności).

Elementy kontekstowe:

- należy jasno określić cele, jakie planuje się osiągnąć poprzez zanonimizowanie zbioru danych, ponieważ odgrywają one ważną rolę w określaniu ryzyka identyfikacji;
- jednocześnie należy uwzględnić wszystkie odpowiednie elementy kontekstowe – np. charakter danych pierwotnych, istniejące mechanizmy kontroli (w tym środki bezpieczeństwa służące ograniczeniu dostępu do zbiorów danych), liczebność próby (cechy ilościowe), dostępność zasobów informacji publicznych (na których mogą opierać się odbiorcy), przewidziane udostępnienie danych osobom trzecim (ograniczone, nieograniczone np. w internecie itp.);
- należy zwrócić uwagę na potencjalnych atakujących, uwzględniając atrakcyjność danych z perspektywy ukierunkowanych ataków (w tym kontekście ponownie najważniejszymi czynnikami będą szczególna ochrona informacji i charakter danych).

Elementy techniczne:

- administratorzy danych powinni ujawnić technikę anonimizacji / połączenie technik, które zastosowano, w szczególności jeżeli planują udostępnić zanonimizowany zbiór danych;
- ze zbioru danych należy usunąć oczywiste (np. rzadkie) atrybuty / *quasi*-identyfikatory;
- jeżeli stosuje się technikę dodawania zakłóceń (w randomizacji), poziom zakłóceń dodanych do zapisów należy określić jako funkcję wartości atrybutu (tj. nie należy dodawać żadnych zakłóceń wykraczających poza skalę), wpływu atrybutów, które mają podlegać ochronie, na osoby, których dane dotyczą, lub rozproszenie zbioru danych;
- w przypadku opierania się na prywatności różnicowej (w randomizacji) należy uwzględnić konieczność monitorowania zapytań, aby wykrywać zapytania naruszające prywatność, ponieważ naruszenia w ramach zapytań mają charakter kumulacyjny;
- jeżeli wdrożono techniki uogólniania, bardzo istotne jest, aby administrator danych nie ograniczał się do jednego kryterium uogólniania nawet w odniesieniu do tego samego atrybutu; oznacza to, że należy wybierać różne poziomy szczegółowości lub różne przedziały czasowe. Wybór kryteriów, które należy stosować, musi zależeć od dystrybucji wartości atrybutu w danej populacji. Nie wszystkie dystrybucje nadają się do uogólniania, tj. w przypadku uogólniania nie można zastosować podejścia uniwersalnego. Należy zapewnić zmienność w ramach klas równoważności; należy na przykład wybrać określony próg na podstawie „elementów kontekstowych", o których mowa powyżej (liczebność próby itp.), i jeżeli próg ten nie zostanie osiągnięty, wówczas należy odrzucić określoną próbę (lub należy określić inne kryterium uogólniania).

ZAŁĄCZNIK
Podręcznik dotyczący technik anonimizacji

A.1. Wprowadzenie

Anonimowość jest różnie interpretowana w całej UE – w niektórych państwach odpowiada anonimowości obliczeniowej (tj. bezpośrednie lub pośrednie zidentyfikowanie jednej z osób, których dane dotyczą, powinno być trudne pod względem obliczeniowym nawet dla administratora danych we współpracy z jakąkolwiek stroną), a w innych anonimowości absolutnej (tj. bezpośrednie lub pośrednie zidentyfikowanie jednej z osób, których dane dotyczą, powinno być niemożliwe nawet dla nawet dla administratora danych we współpracy z jakąkolwiek stroną). W obu przypadkach „anonimizacja" oznacza jednak proces, w wyniku którego dane stają się anonimowe. Różnica polega na tym, co uważa się za dopuszczalny poziom w odniesieniu do ryzyka ponownej identyfikacji.

Można przewidzieć różne przypadki użycia zanonimizowanych danych, np. do celów badań społecznościowych, analiz statystycznych, innowacji usługowej/produktowej. Czasami nawet takie działania o ogólnym celu mogą mieć wpływ na określone osoby, których dane dotyczą, niwecząc rzekomo anonimowy charakter przetworzonych danych. Można podać wiele przykładów – od rozpoczęcia ukierunkowanych inicjatyw marketingowych, po wdrożenie środków publicznych opartych na profilowaniu użytkowników bądź zachowania lub wzorce mobilności[21].

Niestety, poza ogólnymi stwierdzeniami nie istnieje żadna zaawansowana metryka umożliwiająca wcześniejsze oszacowanie czasu lub wysiłku, które są niezbędne do ponownej identyfikacji po przetworzeniu, lub ewentualnie umożliwiająca wybranie najbardziej odpowiedniej procedury, jaką można zastosować, aby zmniejszyć prawdopodobieństwo, że udostępniony zbiór danych będzie się odnosić do zidentyfikowanego zbioru osób, których dane dotyczą.

„Sztuka anonimizacji", jak w literaturze naukowej często określa się te praktyki[22], jest nową gałęzią nauki, która jest jeszcze bardzo słabo rozwinięta, i istnieje wiele praktyk mających na celu obniżenie mocy identyfikacji zbiorów danych; należy jednak jasno stwierdzić, że większość takich praktyk nie zapobiega tworzeniu powiązań między przetworzonymi danymi z osobami, których dane dotyczą. W niektórych okolicznościach identyfikacja zbiorów danych uznanych za anonimowe okazała się bardzo skuteczna, w innych sytuacjach pojawiły się błędne akceptacje.

Zasadniczo istnieją dwa różne podejścia: jedno opiera się na uogólnianiu atrybutu, drugie – na randomizacji. Dokładne zbadanie szczegółów i niuansów tych praktyk da nowe wyobrażenie o kwestii mocy identyfikacji danych i rzuci nowe światło na samo pojęcie danych osobowych.

A.2. „Anonimizacja" przez randomizację

Jeden z wariantów anonimizacji polega na modyfikowaniu faktycznych wartości w celu uniemożliwienia tworzenia powiązań między zanonimizowanymi danymi a wartościami pierwotnymi. Cel ten można osiągnąć za pomocą wielu metod – od dodawania zakłóceń po podstawianie danych (permutacja). Należy zaznaczyć, że usuwanie atrybutu jest równoważne

[21] Na przykład sprawa TomTom w Niderlandach (zob. przykład omówiony w pkt 2.2.3).
[22] Jun Gu,Yuexian Chen,Junning Fu,HuanchunPeng,Xiaojun Ye, „Synthesizing: Art of Anonymization", *Database and Expert Systems Applications*, Lecture Notes in Computer Science –Springer- tom 6261, 2010 r., s. 385–399

z ekstremalną formą randomizacji tego atrybutu (atrybut ten jest w całości objęty zakłóceniami).

W niektórych okolicznościach celem ogólnego przetwarzania jest nie tyle udostępnianie randomizowanego zbioru danych, co raczej przyznanie dostępu do danych za pomocą zapytań. W tym przypadku ryzyko dla osoby, której dane dotyczą, wynika z prawdopodobieństwa, że atakujący będzie w stanie wyciągnąć informacje z szeregu różnych zapytań bez wiedzy administratora danych. W celu zagwarantowania anonimowości osobom fizycznym uwzględnionym w zbiorze danych nie powinno być możliwe stwierdzenie, że osoba, której dane dotyczą, wniosła wkład do zbioru danych, zrywając w ten sposób powiązanie z wszelkiego rodzaju informacjami podstawowymi, jakie atakujący może posiadać.

Dodawanie w razie potrzeby zakłóceń do odpowiedzi na zapytania może dodatkowo ograniczyć ryzyko ponownej identyfikacji. Podejście to, znane także w literaturze jako prywatność różnicowa[23], odchodzi od podejść opisanych wcześniej pod tym względem, że daje podmiotom publikującym dane większą kontrolę nad dostępem do danych w porównaniu z udostępnianiem publicznym. Dodawanie zakłóceń ma dwa główne cele: pierwszym jest ochrona prywatności osób, których dane dotyczą, w zbiorze danych, a drugim zachowanie użyteczności udostępnionych informacji. W szczególności skala zakłóceń musi być proporcjonalna do poziomu zapytań (zbyt wiele zapytań dotyczących poszczególnych osób fizycznych, na które udzielone zostaną zbyt dokładne odpowiedzi, skutkuje zwiększaniem prawdopodobieństwa identyfikacji). Obecnie należy rozpatrywać skuteczne zastosowanie randomizacji w odniesieniu do poszczególnych przypadków, przy czym nie istnieje żadna technika oferująca niezawodną metodę, ponieważ istnieją przykłady wycieków informacji na temat atrybutów osób, których dane dotyczą (niezależnie od tego, czy zostały one uwzględnione w zbiorze danych, czy nie), chociaż administrator danych uważał zbiór danych za randomizowany.

Pomocne może być omówienie konkretnych przykładów w celu wyjaśnienia potencjalnych niepowodzeń randomizacji jako środka zapewniania anonimizacji. Na przykład w kontekście interaktywnego dostępu zapytania uznawane za niezagrażające prywatności mogą stanowić ryzyko dla osób, których dane dotyczą. W praktyce, jeżeli atakujący wie, że podgrupa osób fizycznych S została uwzględniona w zbiorze danych, który zawiera informacje na temat występowania atrybutu A w ramach populacji P, poprzez proste wprowadzenie zapytania w postaci dwóch pytań: „Jak wiele osób w populacji P posiada atrybut A?" i „Jak wiele osób w populacji P poza tymi należącymi do podgrupy S posiada atrybut A?" może istnieć możliwość określenia (przez różnicę) liczby osób w S, które faktycznie posiadają atrybut A – w sposób deterministyczny lub przez wnioskowanie prawdopodobieństwa. W każdym przypadku prywatność osób fizycznych w podgrupie S może zostać poważnie zagrożona, w szczególności w zależności od charakteru atrybutu A.

Można również stwierdzić, że jeżeli osoba, której dane dotyczą, nie jest uwzględniona w zbiorze danych, ale jej związek z danymi w zbiorze danych jest znany, udostępnienie zbioru danych może spowodować zagrożenie dla jej prywatności. Na przykład, jeżeli wiadomo, że „wartość atrybutu A w przypadku celu różni się o ilość X od średniej wartości populacji", poprzez zwykłe poproszenie opiekuna bazy danych o przeprowadzenie operacji niezagrażającej prywatności, jaką jest wyciągnięcie średniej wartości atrybutu A, atakujący może dokładnie wywnioskować dane osobowe określonej osoby, której dane dotyczą.

[23] Cynthia Dwork, „Differential Privacy", [w:] *International Colloquium on Automata, Languages and Programming* (ICALP) 2006 r., s. 1–12

Wprowadzanie pewnych stosunkowych niedokładności do faktycznych wartości w zbiorze danych jest operacją, którą należy odpowiednio zaprojektować. Należy dodać wystarczającą ilość zakłóceń, aby chronić prywatność, ale odpowiednio małą, aby zachować użyteczność danych. Na przykład, jeżeli liczba osób, których dane dotyczą, o szczególnym atrybucie jest bardzo niska lub jeżeli szczególna ochrona atrybutu jest wysoka, lepszym rozwiązaniem jest zgłoszenie zakresu lub zdania ogólnego „mała liczba przypadków, możliwie nawet równa zeru", niż zgłaszanie faktycznej liczby. W ten sposób, nawet jeżeli mechanizm ujawniania zakłóceń jest znany z wyprzedzeniem, prywatność osoby, której dane dotyczą, zostaje zachowana, ponieważ pozostaje pewien stopień niepewności. Z perspektywy użyteczności, jeżeli niedokładność została zaprojektowana prawidłowo, wyniki są nadal użyteczne do celów statystycznych lub podejmowania decyzji.

Randomizacja bazy danych i dostęp do prywatności różnicowej wymagają dalszych rozważań. Po pierwsze, prawidłowa ilość zakłóceń może się znacznie różnić w zależności od kontekstu (rodzaju zapytania, wielkości populacji w bazie danych, charakteru atrybutu i właściwej mu mocy identyfikacji), dlatego nie można przewidzieć żadnego rozwiązania *ad omnia*. Co więcej, z czasem może się zmienić kontekst, a mechanizm interaktywny należy odpowiednio modyfikować. Dostosowywanie poziomu zakłóceń wymaga prześledzenia skumulowanych czynników ryzyka dla prywatności, jakie dla osób, których dane dotyczą, stanowi każdy mechanizm interaktywny. W celu wsparcia administratora danych w określaniu odpowiedniego poziomu zakłóceń, jaki za każdym razem należy wprowadzać do faktycznych danych osobowych, mechanizm dostępu do danych powinien być zatem wyposażony w system powiadamiania o osiągnięciu budżetu „kosztów prywatności", a osoby, których dane dotyczą, mogą być narażone na szczególne ryzyko, jeżeli przedstawiono nowe zapytanie.

Z drugiej strony należy również rozważyć przypadek, w którym wartości atrybutów zostały usunięte (lub zmodyfikowane). Powszechnie stosowanym rozwiązaniem w odniesieniu do nietypowych wartości atrybutów jest usuwanie zbioru danych dotyczących nietypowych osób fizycznych lub usuwanie nietypowych wartości. W tym drugim przypadku istotne jest dopilnowanie, aby sam brak wartości nie stał się elementem służącym do zidentyfikowania osoby, której dane dotyczą.

Kolejnym etapem rozważań jest randomizacja przez zastąpienie atrybutu. Głównym błędnym przekonaniem, jeżeli chodzi o anonimizację, jest równoważenie jej z szyfrowaniem lub kodowaniem kluczem. To błędne przekonanie opiera się na dwóch założeniach, mianowicie: a) że po zastosowaniu szyfrowania względem niektórych atrybutów zapisu w bazie danych (np. nazwiska, adresu, daty urodzenia) lub po zastąpieniu tych atrybutów pozornie randomizowanym ciągiem w wyniku operacji kodowania kluczem takiej jak funkcja skrótu z kluczem taki zapis jest „zanonimizowany"; oraz b) że anonimizacja jest bardziej skuteczna, jeżeli długość klucza jest odpowiednia, a algorytm szyfrowania opiera się na stanie wiedzy naukowej i technicznej. To błędne przekonanie jest powszechne wśród administratorów danych i zasługuje na wyjaśnienie, podobnie jak pseudonimizacja i związane z nią rzekomo mniejsze ryzyko.

Przede wszystkim cele tej techniki są zupełnie inne: szyfrowanie jako praktyka w zakresie bezpieczeństwa ma na celu zapewnienie poufności kanału komunikacji między zidentyfikowanymi stronami (ludźmi, urządzeniami lub częściami oprogramowania / sprzętu komputerowego), aby uniemożliwić podsłuchiwanie lub niezamierzone udostępnienie. Kodowanie kluczem odpowiada semantycznemu przekładowi danych uzależnionemu od klucza tajnego. Z drugiej strony celem anonimizacji jest uniknięcie zidentyfikowania

poszczególnych osób fizycznych poprzez zapobieżenie ukrytemu tworzeniu powiązań między atrybutami a osobą, której dane dotyczą.

Ani samo szyfrowanie, ani samo kodowanie kluczem nie pozwala osiągnąć celu, jakim jest uniemożliwienie zidentyfikowania osoby, której dane dotyczą: ponieważ, przynajmniej w rękach administratora danych, dane pierwotne nadal są dostępne lub nadal istnieje możliwość otrzymania tych danych drogą dedukcji. Samo wdrożenie przekładu semantycznego danych osobowych, jak ma to miejsce w przypadku kodowania kluczem, nie eliminuje możliwości przywrócenia danych do ich pierwotnej struktury poprzez zastosowanie algorytmu w przeciwną stronę lub poprzez ataki siłowe, w zależności od charakteru układów, lub w wyniku naruszenia danych. Szyfrowanie w oparciu o stan wiedzy naukowej i technicznej może zapewnić, aby dane były chronione na wyższym poziomie, tj. szyfrowanie to jest niezrozumiałe dla podmiotów nieznających klucza odszyfrowującego, ale niekoniecznie skutkuje ono anonimizacją. Dopóki klucz lub dane pierwotne będą dostępne (nawet w przypadku zaufanej osoby trzeciej zobowiązanej umową do zapewnienia usługi bezpiecznego depozytu klucza), dopóty nie zostanie wyeliminowana możliwość zidentyfikowania osoby, której dane dotyczą.

Koncentrowanie się wyłącznie na dokładności mechanizmu szyfrowania jako środka zapewniania pewnego stopnia „anonimizacji" zbioru danych jest błędne, ponieważ na ogólne bezpieczeństwo mechanizmu szyfrowania lub funkcji skrótu ma wpływ wiele innych czynników technicznych i organizacyjnych. W literaturze przedstawiono szereg skutecznych ataków, w których całkowicie obchodzi się algorytm, ponieważ wykorzystują one niedoskonałości w opiece nad kluczami (np. istnienie mniej bezpiecznego trybu domyślnego) lub inne czynniki ludzkie (np. słabe hasło do odzyskania klucza). Ponadto wybrany schemat szyfrowania o danej wielkości klucza opracowuje się w celu zapewnienia poufności w odniesieniu do danego okresu (wielkość większości aktualnych kluczy będzie musiała zostać zmieniona około 2020 r.), natomiast proces anonimizacji nie powinien być ograniczony w czasie.

Warto też omówić ograniczenia randomizacji atrybutu (lub zastąpienia i usunięcia), biorąc pod uwagę różne złe przykłady anonimizacji poprzez randomizację, które miały miejsce w ostatnich latach, oraz przyczyny takich niepowodzeń.

Dobrze znanym przypadkiem związanym z udostępnieniem słabo zanonimizowanego zbioru danych jest konkurs Netflix Prize[24]. Jeżeli chodzi o zapis ogólny w bazie danych, w której randomizacji poddano szereg atrybutów odnoszących się do osoby, której dane dotyczą, każdy zapis można nadal podzielić na dwa podzapisy w następujący sposób: {atrybuty randomizowane, atrybuty czytelne}, gdzie atrybutami czytelnymi może być połączenie danych, które ponoć nie są osobowe. Konkretną obserwacją, jaką można poczynić na podstawie zbioru danych z konkursu Netflix Prize, jest spostrzeżenie, że każdy zapis może być reprezentowany przez jeden punkt w przestrzeni wielowymiarowej, w której każdy atrybut czytelny jest współrzędną. Przy zastosowaniu tej techniki każdy zbiór danych można postrzegać jako konstelację punktów w takiej wielowymiarowej przestrzeni, która może charakteryzować się wysokim stopniem rozproszenia, co oznacza, że punkty mogą znajdować się w dużej odległości od siebie. W praktyce mogą być położone tak daleko od siebie, że po podziale przestrzeni na duże regiony, każdy region obejmuje tylko jeden zapis. Nawet wprowadzenie zakłóceń nie przybliża zapisów w wystarczającym stopniu, tak aby dzieliły one ten sam region wielowymiarowy. Przykładowo w eksperymencie dotyczącym Netflix

[24] Arvind Narayanan, Vitaly Shmatikov: „Robust De-anonymization of Large Sparse Datasets". IEEE Symposium on Security and Privacy 2008 r.: s. 111–125

zapisy były wystarczająco nietypowe: w ciągu 14 dni przyznano filmom tylko osiem ocen. Po dodaniu zakłóceń zarówno do ocen, jak i do dat nie odnotowano żadnego nakładania się regionów. Innymi słowy, ten sam wybór tylko ośmiu ocenionych filmów stanowił odzwierciedlenie wyrażonych ocen, które nie były wspólne dla jakichkolwiek dwóch osób, których dane dotyczą, w tej samej bazie danych. Na podstawie tej geometrycznej obserwacji badacze połączyli rzekomo anonimowy zbiór danych Netflix z inną publiczną bazą danych z ocenami filmów (IMDB), znajdując w ten sposób użytkowników, którzy ocenili te same filmy w tym samym przedziale czasowym. Ponieważ w przypadku większości użytkowników występowała wzajemnie jednoznaczna zgodność opinii, informacje dodatkowe wyszukane w bazie danych IMDB mogły zostać importowane do udostępnionego zbioru danych Netflix, dzięki czemu dodano tożsamości do wszystkich rzekomo zanonimizowanych zapisów.

Należy podkreślić, że jest to ogólna właściwość: rezydualna część każdej „randomizowanej" bazy danych nadal ma dużą moc identyfikacji, w zależności od rzadkości połączeń atrybutów rezydualnych. Jest to ostrzeżenie, o którym administratorzy danych powinni zawsze pamiętać przy wyborze randomizacji jako swojego sposobu osiągnięcia docelowej anonimizacji.

Wiele tego typu eksperymentów w zakresie ponownej identyfikacji opierało się na podobnym podejściu polegającym na projekcji dwóch baz danych w tej samej podprzestrzeni. Jest to bardzo skuteczna metoda ponownej identyfikacji, która miała ostatnio wiele zastosowań w różnych obszarach. Na przykład w eksperymencie w zakresie identyfikacji przeprowadzonym w odniesieniu do sieci społecznościowej[25] wykorzystano graf społecznościowy przedstawiający użytkowników opatrzonych pseudonimem w formie oznaczeń. W tym przypadku atrybutami wykorzystanymi w celu identyfikacji były listy kontaktów każdego użytkownika, ponieważ wykazano, że prawdopodobieństwo wystąpienia identycznej listy kontaktów u dwóch użytkowników jest bardzo niskie. Na podstawie tego intuicyjnego założenia ustalono, że podgraf połączeń wewnętrznych bardzo ograniczonej liczby węzłów stanowi możliwą do uzyskania topologiczną cechę charakteryzującą ukrytą w sieci oraz że po zidentyfikowaniu tej podsieci można zidentyfikować znaczną część całej sieci społecznościowej. Celem przedstawienia danych na temat wyników podobnych ataków wykazano, że przy wykorzystaniu mniej niż 10 węzłów (które mogą dać początek milionowi różnych konfiguracji podsieci, z których każda potencjalnie stanowi topologiczną cechę charakteryzującą) sieć społecznościowa, obejmująca ponad 4 mln węzłów opatrzonych pseudonimem i 70 mln powiązań, może być narażona na ataki ponownej identyfikacji, a prywatność dużej liczby powiązań może być narażona na szwank. Należy podkreślić, że to podejście do ponownej identyfikacji nie jest dostosowane tylko do szczególnego kontekstu sieci społecznościowych, ale jest na tyle ogólne, że może zostać potencjalnie przystosowane do innych baz danych, w których rejestruje się stosunki między użytkownikami (np. kontaktów telefonicznych, korespondencji elektronicznej, portali randkowych itp.).

Inny sposób zidentyfikowania rzekomo anonimowego zapisu opiera się na analizie stylu pisania (stylometrii)[26]. Opracowano już wiele algorytmów mających na celu wydobycie metryki z tekstu poddanego analizie składniowej, obejmującej częstotliwości używania określonego słowa, występowania szczególnych gramatycznych wzorów i rodzaju interpunkcji. Wszystkie te właściwości można wykorzystać w celu połączenia rzekomo anonimowego tekstu ze stylem pisania zidentyfikowanego autora. Badacze uzyskali style pisania z ponad 100 000 blogów i obecnie są w stanie automatycznie zidentyfikować autora

[25] L. Backstrom, C. Dwork, i J. M. Kleinberg. „Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography", Proceedings of the 16th International Conference on World Wide Web WWW'07, s.181–190 (2007 r.).

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/

postu z dokładnością sięgającą już 80 %; dokładność tej techniki będzie się dalej zwiększać także przez wykorzystywanie innych sygnałów, takich jak lokalizacja lub inne metadane zawarte w tekście.

Na większą uwagę społeczności badaczy i sektora zdecydowanie zasługuje kwestia mocy identyfikacji z wykorzystaniem semantyki zapisu (tj. nierandomizowanej, rezydualnej części zapisu). Odzyskanie tożsamości dawców DNA, do którego doszło ostatnio (2013 r.)[27], wskazuje, że od czasu głośnej sprawy wycieku danych z AOL (2006 r.), kiedy to udostępniono publicznie bazę danych zawierającą dwadzieścia milionów słów kluczy wpisywanych do wyszukiwarki przez ponad 650 000 użytkowników przez okres 3 miesięcy, poczyniono jedynie bardzo niewielkie postępy. Udostępnienie tej bazy skutkowało zidentyfikowaniem i zlokalizowaniem wielu użytkowników AOL.

Inną rodziną danych, które rzadko anonimizuje się jedynie poprzez usunięcie tożsamości osób, których dane dotyczą, lub poprzez częściowe szyfrowanie niektórych atrybutów, są dane dotyczące lokalizacji. Wzorce mobilności ludzi są lub mogą być na tyle nietypowe, że semantyczna część danych dotyczących lokalizacji (miejsca, w którym osoba, której dane dotyczą, znajdowała się w określonym czasie), nawet bez innych atrybutów, może zdradzić wiele cech osoby, której dane dotyczą[28]. Zostało to wielokrotnie udowodnione w reprezentatywnych badaniach naukowych[29].

W tym względzie konieczne jest ostrzeżenie przed wykorzystywaniem pseudonimów jako sposobu zapewnienia osobom, których dane dotyczą, odpowiedniej ochrony przed wyciekami tożsamości lub atrybutów. Jeżeli pseudonimizacja opiera się na zastąpieniu tożsamości innym unikalnym kodem, założenie, że stanowi ona solidną anonimizację jest naiwne i nie uwzględnia złożoności metod identyfikacji ani różnorodnych kontekstów, w jakich można je stosować.

A.3. „Anonimizacja" poprzez uogólnianie

Podejście oparte na uogólnianiu atrybutu można wyjaśnić przy pomocy prostego przykładu.

Przedmiotem rozważań jest przypadek, w którym administrator danych zdecydował się udostępnić prostą tabelę zawierającą trzy informacje lub atrybuty: numer identyfikacyjny unikalny w odniesieniu do każdego zapisu, identyfikację lokalizacji łączącej osobę, której dane dotyczą, z miejscem jej zamieszkania i identyfikację właściwości, która wskazuje właściwość posiadaną przez osobę, której dane dotyczą; następnie zakłada się, że właściwość ta jest jedną z dwóch różnych wartości wyrażonych ogólnie jako {W1, W2}:

---

[27] Dane genetyczne są niezwykle istotnym przykładem danych szczególnie chronionych, które mogą być narażone na ponowną identyfikację, jeżeli jedynym mechanizmem „anonimizowania" ich jest usuwanie tożsamości dawców. Zob. przykład, o którym mowa w powyższym pkt 2.2.2. Zob. też John Bohannon, „Genealogy Databases Enable Naming of Anonymous DNA Donors", Science, tom 339, nr 6117 (18 stycznia 2013 r.), s. 262.

[28] Kwestia ta była przedmiotem regulacji w niektórych przepisach krajowych. Na przykład we Francji publikowane statystyki dotyczące lokalizacji są anonimizowane poprzez uogólnianie i permutację. Dlatego też INSEE publikuje statystyki, które są uogólnione przez agregację wszystkich danych w obszar wielkości 40 000 metrów kwadratowych. Poziom szczegółowości zbioru danych jest wystarczający, aby zachować użyteczność danych, a permutacje zapobiegają atakom deanonimizacji w obszarach o małym zagęszczeniu. Mówiąc bardziej ogólnie, agregowanie tej rodziny danych i jej permutowanie zapewnia silne gwarancje przeciwdziałające wnioskowaniu i atakom deanonimizacji (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. „Unique in the Crowd: The privacy bounds of human mobility". Nature. 3, 1376 (2013).

Numer	Lokalizacja	Właściwość
#1	Rzym	W1
#2	Madryt	W1
#3	Londyn	W2
#4	Paryż	W1
#5	Barcelona	W1
#6	Mediolan	W2
#7	Nowy Jork	W2
#8	Berlin	W1

Tabela A1. Próba osób, których dane dotyczą, zebrana według lokalizacji i właściwości W1 i W2.

Jeżeli dana osoba, zwana atakującym, z góry wie, że określona osoba, której dane dotyczą, (cel) mieszkająca w Mediolanie jest uwzględniona w tabeli, wówczas po zbadaniu tabeli może się dowiedzieć, że #6 będący jedyną osobą, której dane dotyczą, ze wspomnianą lokalizacją, posiada również właściwość W2.

Ten bardzo podstawowy przykład pokazuje najważniejsze elementy każdej procedury identyfikacyjnej mającej zastosowanie do zbioru danych, który został poddany rzekomemu procesowi anonimizacji. Mianowicie, istnieje atakujący, który (przypadkowo lub świadomie) dysponuje wiedzą podstawową na temat niektórych lub wszystkich osób, których dane dotyczą, w zbiorze danych. Atakujący ma na celu powiązanie tej wiedzy podstawowej z danymi w udostępnionym zbiorze danych, aby uzyskać jaśniejszy obraz charakterystyki tych osób, których dane dotyczą.

Aby dane tworzyły powiązania z każdego rodzaju wiedzą podstawową mniej skutecznie lub w sposób w mniejszym stopniu natychmiastowy, administrator danych może skoncentrować się w szczególności na lokalizacji, zastępując miasto, w którym mieszka osoba, której dane dotyczą, szerszym obszarem, np. państwem. W ten sposób tabela wyglądałaby następująco:

Numer	Lokalizacja	Właściwość
#1	Włochy	W1
#2	Hiszpania	W1
#3	Zjednoczone Królestwo	W2
#4	Francja	W1
#5	Hiszpania	W1
#6	Włochy	W2
#7	Stany Zjednoczone	W2
#8	Niemcy	W1

Tabela A2. Uogólnianie tabeli A1 według narodowości.

Przy tej nowej agregacji danych podstawowa wiedza atakującego na temat zidentyfikowanej osoby, której dane dotyczą, (zgodnie z którą: „obiekt mieszka w Rzymie i jest uwzględniony w tabeli") nie umożliwia wyciągnięcia jakichkolwiek jasnych wniosków na temat jego właściwości: jest tak, ponieważ dwóch Włochów w tabeli ma różne właściwości: odpowiednio W1 i W2. Atakującemu pozostaje 50-procentowa niepewność co do właściwości podmiotu docelowego. Ten prosty przykład pokazuje skutek uogólniania dla praktyki anonimizacji. W istocie, chociaż ten sposób uogólnienia może być skuteczny w zmniejszeniu

o połowę prawdopodobieństwa zidentyfikowania Włocha będącego celem, nie jest on skuteczny w odniesieniu do celów z innych lokalizacji (np. Stanów Zjednoczonych).

Co więcej, atakujący nadal może zdobyć informacje na temat celu hiszpańskiego. Jeżeli wiedzą podstawową jest na przykład to, że „cel mieszka w Madrycie i jest uwzględniony w tabeli" lub że „cel mieszka w Barcelonie i jest uwzględniony w tabeli", atakujący może wywnioskować ze 100-procentową pewnością że cel posiada właściwość W1. Dlatego też uogólnianie nie skutkuje tym samym poziomem prywatności lub oporu przed atakami opartymi na wnioskowaniu dla całej populacji w zbiorze danych.

Zgodnie z powyższym rozumowaniem można się pokusić o stwierdzenie, że silniejsze uogólnianie może być pomocne w uniemożliwianiu tworzenia jakichkolwiek powiązań – na przykład uogólnianie według kontynentu. W ten sposób tabela wyglądałaby następująco:

Numer	Lokalizacja	Właściwość
#1	Europa	W1
#2	Europa	W1
#3	Europa	W2
#4	Europa	W1
#5	Europa	W1
#6	Europa	W2
#7	Ameryka Północna	W2
#8	Europa	W1

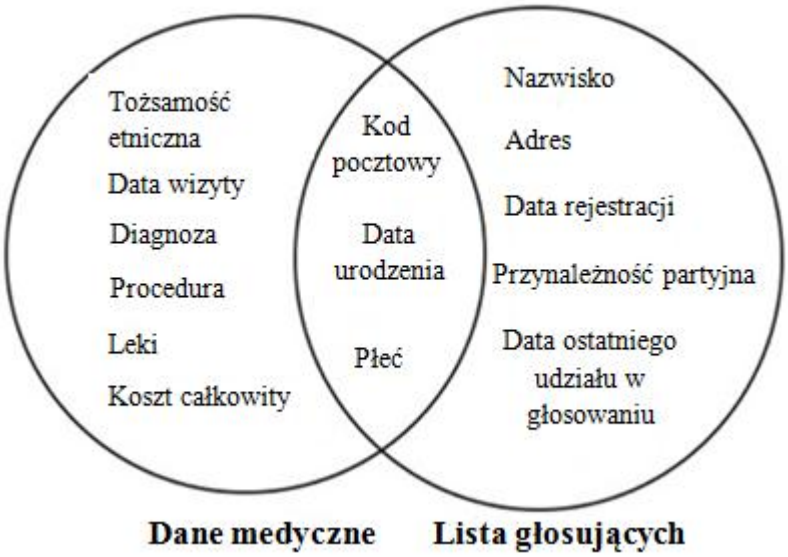
Tabela A3. Uogólnianie tabeli A1 według kontynentu.

Przy tego rodzaju agregacji danych wszystkie osoby, których dane dotyczą, w tabeli, z wyjątkiem tej mieszkającej w Stanach Zjednoczonych, byłyby chronione przed atakami na podstawie tworzenia powiązań i atakami ukierunkowanymi na identyfikację, a wszelkie informacje podstawowe, takie jak „cel mieszka w Madrycie i jest uwzględniony w tabeli" lub „cel mieszka w Mediolanie i jest uwzględniony w tabeli" prowadziłby do pewnego poziomu prawdopodobieństwa w odniesieniu do właściwości mającej zastosowanie do danej osoby, której dane dotyczą (W1 z prawdopodobieństwem 71,4 % i W2 z prawdopodobieństwem 28,6 %), a nie do bezpośredniego powiązania. Ponadto przedmiotowe dalsze uogólnianie odbywa się kosztem wyraźnej i znacznej utraty informacji: tabela nie pozawala na odkrycie potencjalnych korelacji między właściwościami a lokalizacją, mianowicie na ustalenie, czy określona lokalizacja może przyczynić się do zwiększenia prawdopodobieństwa występowania którejkolwiek z dwóch właściwości, ponieważ skutkuje ona tylko tak zwaną dystrybucją „marginalną", czyli prawdopodobieństwem absolutnym wystąpienia właściwości W1 i W2 w całej populacji (odpowiednio 62,5 % i 37,5 % w analizowanym przykładzie) i na każdym kontynencie (odpowiednio, jak wspomniano: 71,4 % i 28,6 % w Europie oraz 100 % i 0 % w Ameryce Północnej).

Przykład ten pokazuje również, że stosowanie uogólniania wpływa na praktyczną użyteczność danych. Niektóre narzędzia projektowania są obecnie dostępne w celu wcześniejszego (tj. przed udostępnieniem zbioru danych) ustalenia, jaki poziom uogólnienia atrybutu jest najbardziej odpowiedni, aby ograniczyć ryzyko zidentyfikowania w tabeli osób, których dane dotyczą, bez nadmiernego wpływu na użyteczność udostępnionych danych.

k-anonimizacja

Dążenie do uniemożliwienia ataków opartych na tworzeniu powiązań, na podstawie uogólniania atrybutów, jest znane jako k-anonimizacja. Praktyka ta wywodzi się z eksperymentu w zakresie ponownej identyfikacji przeprowadzonego w późnych latach 90. XX wieku, w ramach którego prywatne przedsiębiorstwo amerykańskie prowadzące działalność w sektorze zdrowia udostępniło publicznie rzekomo zanonimizowany zbiór danych. Anonimizacja ta polegała na usuwaniu nazwisk osób, których dane dotyczą, ale zbiór danych nadal zawierał dane dotyczące zdrowia i inne atrybuty, takie jak kod pocztowy (lokalizacja, w której osoby te mieszkały), płeć i pełna data urodzenia. Te same trzy atrybuty {kod pocztowy, płeć, pełna data urodzenia} zostały również uwzględnione w innych publicznie dostępnych rejestrach (np. na liście wyborców) i dlatego pracownik naukowy mógł je wykorzystać w celu powiązania tożsamości określonej osoby, której dane dotyczą, z atrybutami w udostępnionym zbiorze danych. Atakujący (badacz) mógł posiadać następującą wiedzę podstawową: „Wiem, że osoba, której dane dotyczą, znajdująca się na liście wyborców zawierającej trzy określone atrybuty {kod pocztowy, płeć, pełna data urodzenia} jest niepowtarzalna. W udostępnionym zbiorze danych istnieje zapis z tymi trzema atrybutami". Jak zaobserwowano w praktyce[30], przeważająca większość (ponad 80 %) osób, których dane dotyczą, w rejestrze publicznym wykorzystanym w tym eksperymencie była jednoznacznie powiązana z trzema określonymi atrybutami, co umożliwiło identyfikację. Dlatego też w tym przypadku dane nie zostały odpowiednio zanonimizowane.



Rysunek A1. Ponowna identyfikacja przez tworzenie powiązań.

Stwierdzono, że w celu ograniczenia skuteczności podobnych ataków opartych na tworzeniu powiązań administratorzy danych powinni najpierw zbadać zbiór danych i pogrupować te atrybuty, które mogą prawdopodobnie zostać wykorzystane przez atakującego do powiązania udostępnionej tabeli z innym dodatkowym źródłem; każda grupa powinna obejmować przynajmniej *k* identycznych kombinacji uogólnionych atrybutów (tj. powinna reprezentować klasę równoważności atrybutów). Zbiory danych należy następnie udostępnić dopiero po podzieleniu ich na takie homogeniczne grupy. Atrybuty wybrane do uogólnienia są znane w literaturze jako *quasi*-identyfikatory, ponieważ znajomość tych atrybutów w formie czytelnej wiązałaby się z natychmiastowym zidentyfikowaniem osób, których dane dotyczą.

[30] L. Sweeney. „Weaving Technology and Policy Together to Maintain Confidentiality". Journal of Law, Medicine &Ethics, 25, nr 2 i 3 (1997): 98–110.

Wiele eksperymentów w zakresie identyfikacji pokazało wady słabo zaprojektowanych tabel poddanych k-anonimizacji. Może tak się stać, na przykład, ponieważ pozostałe atrybuty w klasie równoważności są identyczne (jak ma to miejsce w przypadku klasy równoważności osób, których dane dotyczą, będących Hiszpanami w przykładzie w tabeli A2) lub ich dystrybucja jest bardzo niezrównoważona, z dużą przewagą określonego atrybutu, lub ponieważ liczba zapisów w klasie równoważności jest bardzo mała, pozwalając w obu przypadkach na wnioskowanie na podstawie prawdopodobieństwa, lub ponieważ nie istnieje żadna znacząca różnica „semantyczna" między atrybutami czytelnymi w klasach równoważności (np. ilościowy pomiar takich atrybutów może być faktycznie różny, ale numerycznie zbliżony, lub mogą one należeć do zakresu semantycznie podobnych atrybutów, np. tego samego zakresu ryzyka kredytowego lub tej samej rodziny patologii), przez co nadal istnieje możliwość wycieku ze zbioru danych dużej ilości informacji na temat osób, których dane dotyczą, w drodze ataków opartych na tworzeniu powiązań[31]. Należy przy tym zauważyć, że w każdym przypadku, gdy dane są rozproszone (na przykład na danym obszarze geograficznym dana właściwość występuje nielicznie) i w ramach pierwszej agregacji nie ma możliwości pogrupowania danych z wystarczającą liczbą wystąpień różnych właściwości (na przykład nadal w danym obszarze geograficznym można zlokalizować małą liczbę wystąpień kilku właściwości), konieczna jest dalsza agregacja atrybutów, aby osiągnąć docelową anonimizację.

l-dywersyfikacja

Na podstawie powyższych obserwacji na przestrzeni lat zaproponowano różne warianty k-anonimizacji; opracowywano też określone kryteria projektowania ukierunkowane na udoskonalenie praktyki anonimizacji przez uogólnianie, co miało na celu ograniczenie ryzyka ataków opartych na tworzenia powiązań. Opierają się one na probabilistycznych właściwościach zbiorów danych. W szczególności dodaje się dalsze ograniczenie, tj., że każdy atrybut w klasie równoważności występuje co najmniej *l* razy, przez co atakujący nigdy nie ma dużej pewności co do atrybutów, nawet jeżeli dysponuje wiedzą podstawową na temat określonej osoby, której dane dotyczą. Jest to równoznaczne ze stwierdzeniem, że zbiór danych (lub przedział) powinien posiadać minimalną liczbę wystąpień wybranej właściwości: ta metoda może ograniczyć ryzyko ponownej identyfikacji. Jest to celem praktyki anonimizacji w formie l-dywersyfikacji. Przykład tej praktyki przedstawiono w tabelach A4 (dane pierwotne) i A5 (wynik przetwarzania). Jak się okazuje, dzięki odpowiedniemu zaprojektowaniu lokalizacji i wieku osób fizycznych uwzględnionych w tabeli A4 uogólnianie atrybutów skutkuje znacznym wzrostem niepewności co do faktycznych atrybutów każdej osoby, której dane dotyczą, uczestniczącej w badaniu. Na przykład, nawet jeżeli atakujący wie, że osoba, której dane dotyczą, należy do pierwszej klasy równoważności, nie może dalej sprawdzić, czy osoba ta posiada właściwość X, Y czy Z, ponieważ w tej klasie (i w każdej innej klasie równoważności) istnieje co najmniej jeden zapis wykazujący takie cechy.

[31] Należy podkreślić, że korelacje można również ustalić po pogrupowaniu zapisów danych według atrybutów. Jeżeli administrator danych zna rodzaje korelacji, które chce zweryfikować, może wybrać najbardziej odpowiednie atrybuty. Na przykład wyniki badania PEW nie są przedmiotem bardzo szczegółowych ataków na podstawie wnioskowania i nadal są bardzo użyteczne w znajdowaniu korelacji między demografią a zainteresowaniami. (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

Numer porządkowy	Lokalizacja	Wiek	Właściwość
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabela A4. Tabela przedstawiająca osoby fizyczne pogrupowane według lokalizacji, wieku i trzech właściwości X, Y i Z.

Numer porządkowy	Lokalizacja	Wiek	Właściwość
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tabela A5. Przykład wersji tabeli A4 poddanej l-dywersyfikacji.

t-bliskość:

W odniesieniu do szczególnego przypadku atrybutów w ramach przedziału, które są nierównomiernie rozłożone lub należą do małego zakresu wartości lub znaczeń semantycznych, stosuje się podejście znane jako t-bliskość. Jest ono dalszym udoskonaleniem anonimizacji poprzez uogólnianie i polega na praktyce porządkowania danych w celu osiągnięcia klas równoważności, które w możliwie największym stopniu odzwierciedlają początkową dystrybucję atrybutów w pierwotnym zbiorze danych. W tym celu stosuje się dwuetapową procedurę, która zasadniczo wygląda następująco: tabela A6 stanowi pierwotny zbiór danych obejmujących czytelne zapisy osób, których dane dotyczą, pogrupowane według lokalizacji, wieku, wynagrodzenia i dwóch rodzin semantycznie podobnych właściwości, odpowiednio (X1, X2, X3) i (Y1, Y2, Y3) (np. podobne klasy ryzyka kredytowego, podobne choroby). Najpierw tabelę poddaje się *l-dywersyfikacji*, przy czym *l*=1 (tabela A7), poprzez grupowanie zapisów w semantycznie podobne klasy równoważności i poprzez niedokładną anonimizację docelową; następnie przetwarza się ją w celu uzyskania t-bliskości (tabela A8) i wyższej zmienności w każdym przedziale. W praktyce na drugim etapie każda klasa równoważności obejmuje zapisy z obu rodzin właściwości. Warto zauważyć, że lokalizacja i wiek mają różne poziomy szczegółowości na różnych etapach przetwarzania: oznacza to, że każdy atrybut może wymagać innych kryteriów uogólniania w celu uzyskania docelowej anonimizacji, a to z kolei wymaga od określonego projektowania i odpowiedniego obciążenia obliczeniowego po stronie administratorów danych.

Numer porządkowy	Lokalizacja	Wiek	Wynagrodzenie	Właściwość
1	1127	29	30 tys.	X1
2	1112	22	32 tys.	X2
3	1128	27	35 tys.	X3
4	1215	43	50 tys.	X2
5	1219	52	120 tys.	Y1
6	1216	47	60 tys.	Y2
7	1115	30	55 tys.	Y2
8	1123	36	100 tys.	Y3
9	1117	32	110 tys.	X3

Tabela A6. Tabela przedstawia osoby fizyczne pogrupowane według lokalizacji, wieku, wynagrodzenia i dwóch rodzin właściwości.

Numer porządkowy	Lokalizacja	Wiek	Wynagrodzenie	Właściwość
1	11**	2*	30 tys.	X1
2	11**	2*	32 tys.	X2
3	11**	2*	35 tys.	X3
4	121*	>40	50 tys.	X2
5	121*	>40	120 tys.	Y1
6	121*	>40	60 tys.	Y2
7	11**	3*	55 tys.	Y2
8	11**	3*	100 tys.	Y3
9	11**	3*	110 tys.	X3

Tabela A7. Wersja tabeli A6 poddana *l*-dywersyfikacji.

Numer porządkowy	Lokalizacja	Wiek	Wynagrodzenie	Właściwość
1	112*	<40	30 tys.	X1
3	112*	<40	35 tys.	X3
8	112*	<40	100 tys.	Y3
4	121*	>40	50 tys.	X2
5	121*	>40	120 tys.	Y1
6	121*	>40	60 tys.	Y2
2	111*	<40	32 tys.	X2
7	111*	<40	55 tys.	Y2
9	111*	<40	110 tys.	X3

Tabela A8. Wersja tabeli A6 poddana t-bliskości.

Należy jasno stwierdzić, że cel uogólniania atrybutów osób, których dane dotyczą, przy pomocy tak zaawansowanych sposobów można czasem osiągnąć wyłącznie w odniesieniu do małej liczby zapisów, a nie w odniesieniu do wszystkich tych zapisów. Dobre praktyki powinny zapewniać, aby każda klasa równoważności obejmowała wiele osób fizycznych i aby nie pozostawała jakakolwiek możliwość przeprowadzenia ataku opartego na wnioskowaniu. W każdym przypadku podejście to wymaga dogłębnej oceny dostępnych danych przez administratorów danych razem z łączoną oceną różnych rozwiązań alternatywnych (na przykład różny zakres amplitud, różne lokalizacje lub poziom szczegółowości wieku itp.). Innymi słowy, anonimizacja poprzez uogólnianie nie może być rezultatem podjęcia przez administratorów danych tylko ogólnej i pojedynczej próby zastąpienia analitycznych wartości atrybutów w zapisie zakresami, ponieważ konieczne są bardziej szczegółowe podejścia ilościowe – takie jak ocena entropii atrybutów w każdym przedziale lub pomiar odległości między pierwotną dystrybucją atrybutów a dystrybucją w każdej klasie równoważności.