Information Modelling and Knowledge Bases XXXIII
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA210486

Content Analysis of Medical and Health Apps' Privacy Policies

Boštjan BRUMEN
University of Maribor (www.um.si), Faculty of Electrical Engineering and Computer science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract. Privacy is a fundamental human right and is widely end extensively protected in the western industrialized world. The recent advances in technologies, especially in the use of applications developed and designed for mobile devices, have led to the rise of its abuse on one hand and a higher awareness of the importance of privacy on the other side. Legal texts protecting privacy have attempted to rectify some of the problems, but the ecosystem giants and mobile apps developers adapted. In this paper, we analyze which data mobile apps developers are collecting. We have focused on a sample of apps in the medical and health field. The research was done using collocations analysis. A relationship between a base word and its collocative partners was sought. The initial visual results have led us to more detailed studies that unveiled some worrying patterns. Namely, applications are collect data about the users and their activities, but also about their family members, medical diagnoses, treatments, and alike, going well beyond the "need to function" / functionality threshold.

Keywords. privacy, GDPR, collocation, apps, similarity, medical, health, fitness

1. Introduction

Mobile phones have become a part of our everyday life and a potent multifunctional tool [1]. Mobile technologies have changed our habits and behaviors drastically. By the end of 2020, mobile internet traffic was around 50 % of total web traffic, with Africa and Asia reaching about 60 % [2]. Daily routine previously conducted in an offline world has shifted online and increasingly to mobile devices. Reading mail turned to read email, browsing newspapers, checking the news and journals on the phone, and going shopping by car has moved to sites like Amazon, eBay, Rakuten, Apple, Aliexpress, and others. Visiting friends turned into retweeting their tweets and liking their posts. A smartphone has taken over many users' central functions without a need for profound technological and science skills.

Because of the ease of use and wide availability of services and apps, users increasingly use smartphones and apps. Developers are developing them to fulfill the users' needs. However, there are costs associated with developing applications. There are several business models apps developers can use, such as "Free," "Freemium," "Subscription," "Paid," and "Paymium" [3]. The most popular business model to monetize an app is Free / Freemium (i.e., advertising) because of its easy implementation and wide acceptance by mobile application users [4]. Over half of apps contain advertising [5].

For ads to be effective, they need to "fit" or "match" the target audience (person) as closely as possible. The matching of users' preferences and ads is done via real-time bidding (RTB) coupled with the app's data. In the bidding process, the ad technology companies auction off ad space in the apps made available by developers by sharing sensitive users' data collected by the app, such as location, device ID, cookies, browsing history, and any other collected data. The sharing is done with many different companies involved in a highly complex process [6].

Sharing of vast amounts of personal data opens many privacy-related questions. Namely, privacy is a human right and is protected in the western world. However, microtargeting the individuals with ads (and other activities) based on their personal data is doable legally, without disclosure and (proper!) informed consent, completely bypassing laws and regulations [7], or at least their intentions to protect the privacy.

Privacy protection is increasingly important in the medical domain [8] as medical data are extremely sensitive and need special protection. However, users using medical apps agree to the terms and conditions and associated privacy policies set by application developers. Users typically do not read privacy policies and hence do not know what they are sharing with the application developers and due to the prevailing "free" business model also with (too) many other companies.

In this paper, the research question is which types of personal data medical apps are collecting. We will answer this question using linguistic analysis of privacy policies and visualization techniques to present the findings.

The rest of the paper is organized as follows. Section 2 presents the literature review dealing with medical apps, data collection, and privacy. In Section 3, we describe our research method and deliver the results. In Section 4, we conclude the paper with final remarks.

2. Literature review

First, we give a brief definition and a description of privacy, followed by a description of the advertisement ecosystem in the mobile apps world.

Privacy has many aspects [9, 10], including :

- informational privacy (e.g., confidentiality, anonymity, secrecy, and data security);
- physical privacy (e.g., modesty and bodily integrity);
- associational privacy (e.g., intimate sharing of personal events);
- proprietary privacy (e.g., self-ownership and control over intangibles such as personal identifiers and genetic data, and tangibles, e.g., ownership of objects); and
- decisional privacy (e.g., autonomy and choice in decision-making).

In this paper, we primarily deal with informational and intangibles proprietary privacy. Privacy belongs to fundamental human rights and has a special place in legal texts. It is explicitly stated under Article 12 of the 1948 Universal Declaration of Human Rights and protected by the 1st, 3rd, 4th, and 5th Amendments of the U.S. Constitution [11]. In European Union, it is protected by Article 8(1) of the Charter of Fundamental Rights of the European Union and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), and several national constitutions [11].

The implementational part of privacy protection is done on the lower level of the hierarchy of laws.

In the European Union, it is protected by General Data Protection Regulation (GDPR) directive [12] – the Regulation (E.U.) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. In the United States, there is no single (federal level) data or privacy protection law [13].

California’s California Consumer Privacy Act (CCPA) is the latest and strictest state law, effective on January 1st, 2020. It was amended and extended by the California Privacy Rights Act and will take effect on January 1st, 2023. The law applies to companies that do business in California. CCPA gives users the right to opt out of the sale of their personal information. Interestingly, there are exceptions to the definition of “sale,” such that not all personal information transfers are sales, e.g., transferring personal information to a service provider is not a “sale” [14].

The above laws are exemplary because they limit or prohibit the unauthorized collection and use (selling) of personal data. They also require the data controllers (collectors) to indicate what type of personal information is being collected and why with special rights vested upon users. Typically, companies disclose their privacy practices in the “Privacy Policy.”

However, privacy policies are hard to read lengthy legal texts in practice, and users do not read them. For example, the Norwegian Consumer Council has conducted an “AppFail” campaign. They have downloaded the terms of service and privacy policies for a set of typical apps. Together they exceed the New Testament in length – and would take more than 24 hours to read out loud [15]. The campaign highlighted the absurd length of these agreements. To use apps, the users often need to waive fundamental privacy rights and agree that apps track them, and personally identifiable data can be resold [16].

By using the apps, users consent to data collection and other practices as described in the privacy policies, not actually knowing what and how is being collected and processed.

On the other hand, tech firms need personal data residing in apps to serve users with targeted ads. Marketing campaigns depend on the successfully placed ads [6]. Data need to be shared (legally) with advertisers, and they are asked to bid on an individual ad space within an application.

The bidding process is a multi-level, complex, automated placement of third-party ads known as real-time bidding (RTB) [17]. The partial data flow is depicted in Figure 1 (source: [18]).

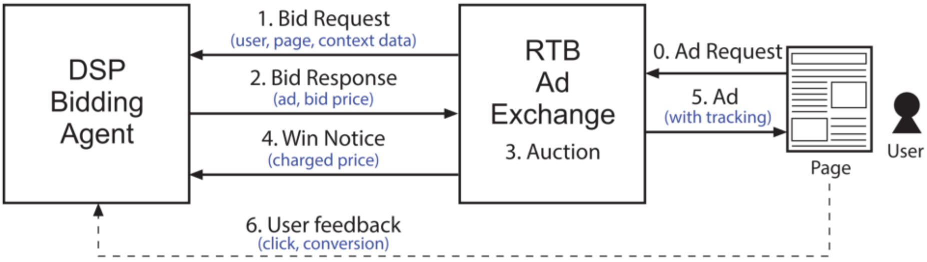


Figure 1: Flow of data and money in an RTB Ad Exchange (Source/Figure from: [18])

In the RTB process, there are several layers of companies involved. It is facilitated by mobile supply-side platforms (SSPs), providing the developers the necessary tools (Software Development Kits, SDKs) that developers build into their apps; the SDKs connect apps to the exchanges. Inside an app, the SSP collects information and enriches it with its own data about the user. The enriched data is sent to ad exchanges that contact demand-side platforms and DSPs to bid for the ad place. DSPs do data matching and possibly bid for the ad space. If the bid is won, a DSP serves a user with an advertisement. The advertiser pays to the DSPs, they pay to the RTB bid exchange (which collaborates closely with SSPs), and the developer gets paid.

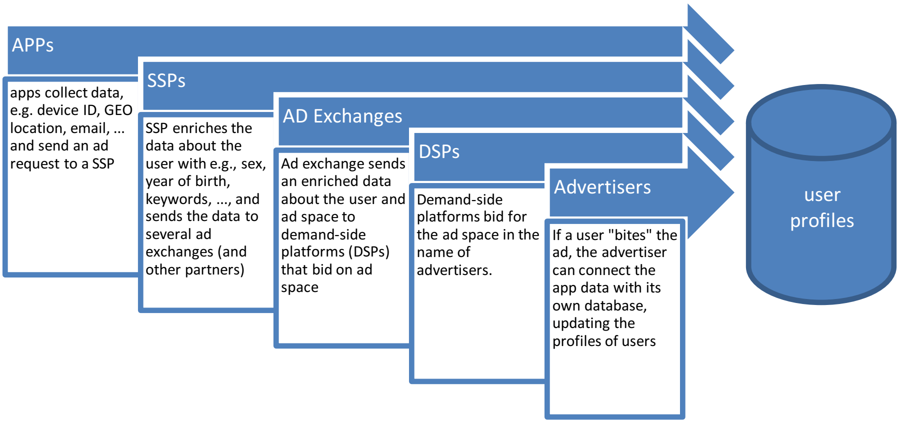

APPs
apps collect data, e.g. device ID, GEO location, email, ... and send an ad request to a SSP
SSPs
SSP enriches the data about the user with e.g., sex, year of birth, keywords, ..., and sends the data to several ad exchanges (and other partners)
AD Exchanges
Ad exchange sends an enriched data about the user and ad space to demand-side platforms (DSPs) that bid on ad space
DSPs
Demand-side platforms bid for the ad space in the name of advertisers.
Advertisers
If a user "bites" the ad, the advertiser can connect the app data with its own database, updating the profiles of users
user profiles

Figure 2: Process flow in a typical ad exchange system

The process of a typical ad exchange system is presented in Figure 2. There are several partners (ad exchanges) competing for advertisement space. Table 1 lists an excerpt of exchanges participating with AdMob, a Google-owned company [19].

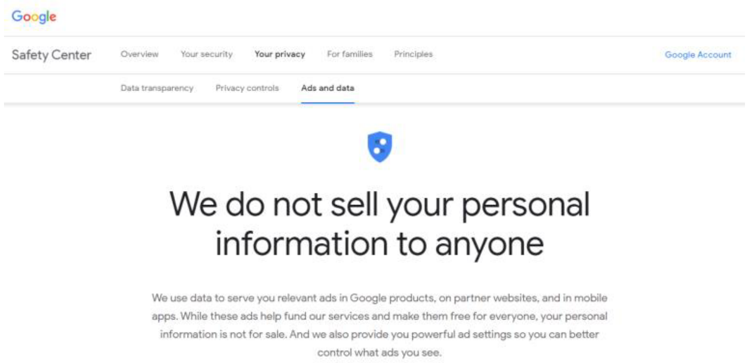
An RTB system will invade privacy in two ways. Firstly, before the RTB process begins, a myriad of companies have already tracked users, collected their personal information online and offline, and combined them into lengthy user profiles. Again, during the RTB process, a set of companies use these previously acquired profiles to decide how much to pay for the ad space. Secondly, due to the advertisement being displayed, the user leaves some traces behind (e.g., data about click, how long an ad was shown, how the users tracked the ad using eye-tracking algorithms, conversion of clicks to purchases, and many others). The companies involved in the RTB process collect these data and update their profiles, knowing even more about users. The newly updated profiles are feeding the future RTBs. RTB is both a cause of tracking and a means of tracking of personal information [20].

Table 1: A partial list of partner RTBs

Open Bidding Partner*	Desktop Web	Mobile Web	Mobile App iOS/Android/Interstitial
Ad Generation	✓	✓	✓
AerServ	✓	✓	✓
Fluct	✓	✓	✓
Improve Digital	✓	✓	✓
Index Exchange	✓	✓	✓
Magnite	✓	✓	✓
Media.net	✓	✓	✓
MobFox	✗	✓	✓
OpenX	✓	✓	✓
PubMatic	✓	✓	✓
ShareThrough	✓	✓	✗
Smaato	✓	✓	✓
Smart Adserver	✓	✓	✓
Sonobi	✓	✓	✓
Sovrn	✓	✓	✓
SpotX	✓	✓	✓
TripleLift	✓	✓	✓
UnrulyX	✓	✓	✓
Verizon Media	✓	✓	✓
Yieldmo	✓	✓	✓
YieldOne	✓	✓	✓

** Some exchanges participate in Open Bidding but have requested not to be listed in this table.*

Google controls massive portions of nearly every level of the real-time bidding ecosystem. It is the owner of DoubleClick, an ad network, and AdMob, the largest ad server for the app market [17]. Tech giants send data to advertisers, and advertisers pay to companies. Nevertheless, Google claims it is not selling personal information, as depicted in Figure 3.

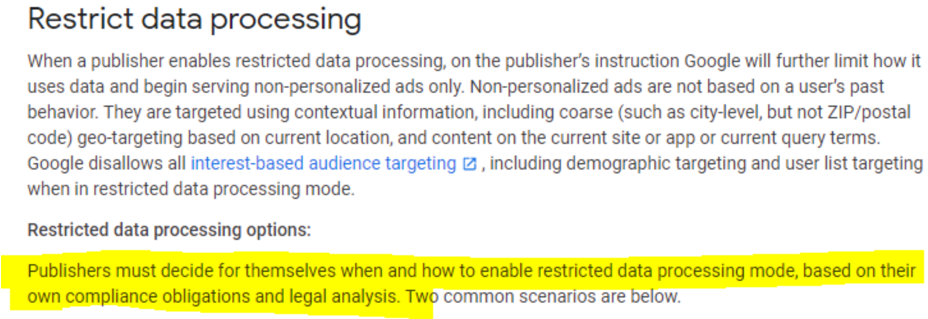


We do not sell your personal information to anyone

We use data to serve you relevant ads in Google products, on partner websites, and in mobile apps. While these ads help fund our services and make them free for everyone, your personal information is not for sale. And we also provide you powerful ad settings so you can better control what ads you see.

Figure 3: Google's claim about selling personal information (https://safety.google/privacy/ads-and-data/, the screenshot was taken 2021-02-05)

It acknowledges that a "sale" is occurring somewhere in this process. Google just insists that they are not selling data, thus breaching the laws [17]. Although facilitating the RTB process, Google places the responsibility for compliance with the rules upon apps' publishers [21], see Figure 4.



Restrict data processing

When a publisher enables restricted data processing, on the publisher's instruction Google will further limit how it uses data and begin serving non-personalized ads only. Non-personalized ads are not based on a user's past behavior. They are targeted using contextual information, including coarse (such as city-level, but not ZIP/postal code) geo-targeting based on current location, and content on the current site or app or current query terms. Google disallows all interest-based audience targeting ↗, including demographic targeting and user list targeting when in restricted data processing mode.

Restricted data processing options:

Publishers must decide for themselves when and how to enable restricted data processing mode, based on their own compliance obligations and legal analysis. Two common scenarios are below.

Figure 4: Google's policy on compliance obligations of apps publishers involved in the advertisement business

As described above, a successful advertisement campaign depends on "good" data being collected at the apps' side. A vicious cycle starts with the good quality and quantity of data an app developer makes available in the RTB process. Based on "good" data, "good" ads will be displayed to get the user's attention, finally converging to a purchase.

For publishers to get lots of personal data, they need to get users' consent to be rules compliant. The consent is typically hidden in privacy policies. The whole advertisement ecosystem relies on users neither reading privacy policies nor understanding or reacting adequately.

The privacy policies are written in a natural language. A collocation analysis of privacy policies, specifically their parts on personal data collection, can be conducted to extract interesting patterns. A collocation is a set of terms that co-occur more often than would be expected by chance. Collocation can be looked at from three perspectives [22]: a statistical view (co-occurrence), a construction view (a correlation between a lexeme and a lexical-grammatical pattern, or as a relation between a base and its collocative partners), and an expression view (collocations as units of expression). In the present research, we use the statistical and the construction view to extract interesting patterns in the form of keywords [23, 24].

In the following sections, we present a method of analyzing the privacy policies with respect to what medical apps are collecting personal data and the analysis results.

3. Co-occurrence analysis of privacy policies

This section presents the results of the co-occurrence analysis of a sample of free health and medical apps' privacy policies.

3.1. Data collection

First, we scraped the Google Play store for the field "mininstalls" using the public "google-play-scraper" from GitHub repository (https://github.com/JoMingyu/google-play-scraper). From the category "medical" and "health & fitness," we selected the 50 most installed and free ones (number of installs as reported by Google in Google Play Store) in each group. Next, we downloaded the texts of the privacy policies from the URL obtained via scraping from the Google Play store.

The exclusion criteria were that the app should have a privacy policy in the first place and be written in English. Out of 2x50, we ended up with a sample of n=32 privacy policies in “medical” and m=30 in “health & fitness.”

3.2. Data preparation and processing

We manually trimmed the irrelevant text of the policies and kept only the parts describing which private data an app is collecting. Next, the documents were tokenized, and part-of-speech details were added (e.g., the word “you're” was tokenized into “you” and “are”). Punctuations, stop words, and short words (with two characters or less) were removed, lemma details were added, and all text was converted to lowercase – see Listing 1.

```
% begin of data reading and processing
fileName = 'data2.xlsx';                            % file name to read
data = readtable(fileName,'TextType','string');     % read data from text file
                                                    % as a string
head(data);                                         % get header line
textData = data.Description;                        % relevant data is in
                                                    % “description” column

docs = tokenizedDocument(textData);                 % tokenize the data
docs = addPartOfSpeechDetails(docs);                % add part of speech
docs = erasePunctuation(docs);                      % erase punctuations
docs = removeStopWords(docs);                       % remove stopwords
docs = removeShortWords(docs,2);                    % remove short words
docs = removeEmptyDocuments(docs);                  % remove empty lines
docs = addLemmaDetails(docs);                       % add lemmas
docs = lower(docs);                                 % convert to lowercase
% end of data reading and processing
```

Listing 1: data processing source file in MatLab®

After the data preparation, we processed and analyzed the data using MathLab® R2020b, using co-occurrence analysis with Text Analysis Toolbox [25] – see Listing 2.

```
% begin of co-occurrence analysis with
span = 6;                                           % length of connecting arcs
nCooc = 8;                                          % number of co-occurrences
mode = 'backward';                                  % mode of co-occurrence building

nKeywords = 6;                                      % start with 6 words for the
                                                    % general picture

% for detailed pictures, limit visualization to the following stem words:
% nKeywords = ["password"; "fitness"; "family"; "sex"];
% nKeywords = ["location"; "service"; "device"; "browser"]; nCooc=6;
% nKeywords = ["activity"; "heart"; "weight"; "height"; "body"];
% nKeywords = ["route"; "exercise"; "photos"; "gender"];


% build co-occurrence table (function from Text Analysis Toolbox):
COTable   = createCOTable(docs, span, nKeywords, nCooc, mode);

% build co-occurrence network from the table (function from Text Analysis Toolbox):
CONetwork = createCONetwork(COTable, 'MI');

% visualize the network:
visualizeCO(CONetwork, 'NodeColor', [0.5 0.5 0.8], 'FontSize', 12, 'MarkerSize', 55,
'EdgeColor', [0 0.8 0.4]); %, 'EdgeWidth',8);
```

Listing 2: data processing source file in MatLab®

The word cloud was also created using MatLab®, as in Listing 3. The first part, data reading & preparation, is the same as in Listing 1 and thus omitted in Listing 3.

```
% begin word cloud
cleanedBag = bagOfWords(docs);                      % create bag of words
cleanedBag = removeInfrequentWords(cleanedBag,2);   % remove rare words
tbl = topkwords(cleanedBag,63);                     % use top words only
colors = rand(63,3);                                % set random colors
figure
wordcloud (tbl, "Word", "Count", 'Color', colors);  % create figure with word cloud
% end word cloud
```

Listing 3: word cloud file in MatLab®

3.3. Results

Firstly, we calculated single words' frequencies and visualized the results using the word cloud for the medical category (Figure 5) and the health & fitness (Figure 6).

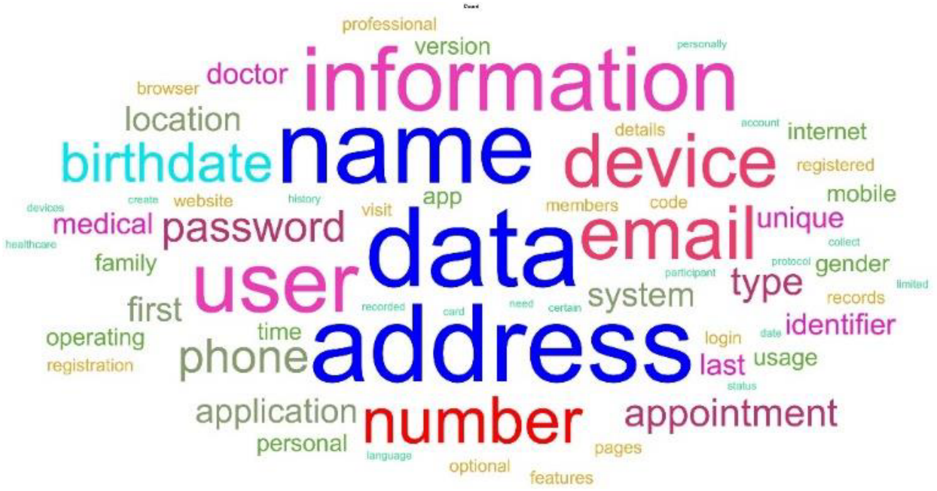


Figure 5: Word cloud of data collection parts of privacy policies – category "medical"

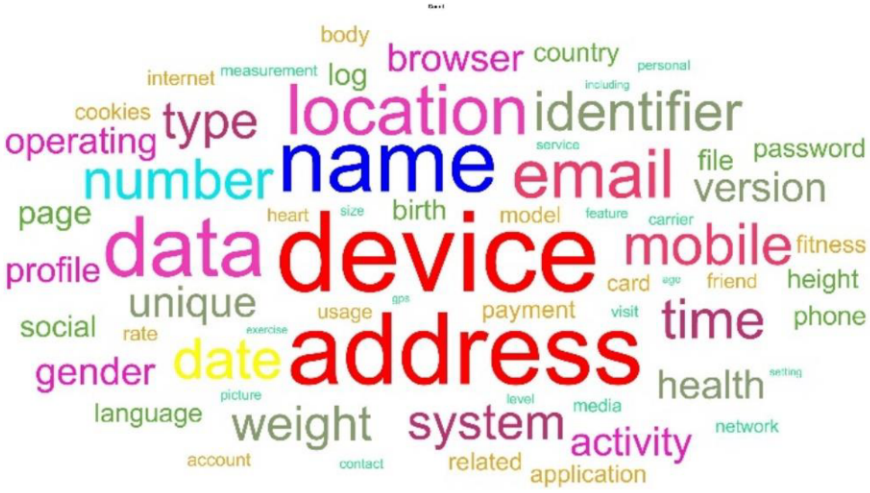


Figure 6: Word cloud of data collection parts of privacy policies – category "health & fitness"

The visualization of individual words' frequencies shows that mostly sought-after pieces of information are the user's email and physical address, phone number, birth date, (geo)location, and data related to browsing activities. Interestingly, from the first glimpse, there are no significant differences between the groups.

Single-word analysis gives only a glimpse into what words are most frequently used in privacy policies.

Here, we already see interesting keywords popping up, such as "name," "password," and "address" in both categories, and "medical," "family," "member," "doctor" in the group "medical," and "body," "gender," "activity," "exercise" in the group "health & fitness."

We will further analyze these words using a targeted co-occurrence analysis.

Secondly, to put words in context, the co-occurrence analysis was conducted. The results were visualized and manually inspected.

3.3.1. Co-occurrence analysis of the "Medical" privacy policies

The visualization result of the co-occurrence analysis of the keywords of privacy policies in the category "medical" is presented in Figure 7.

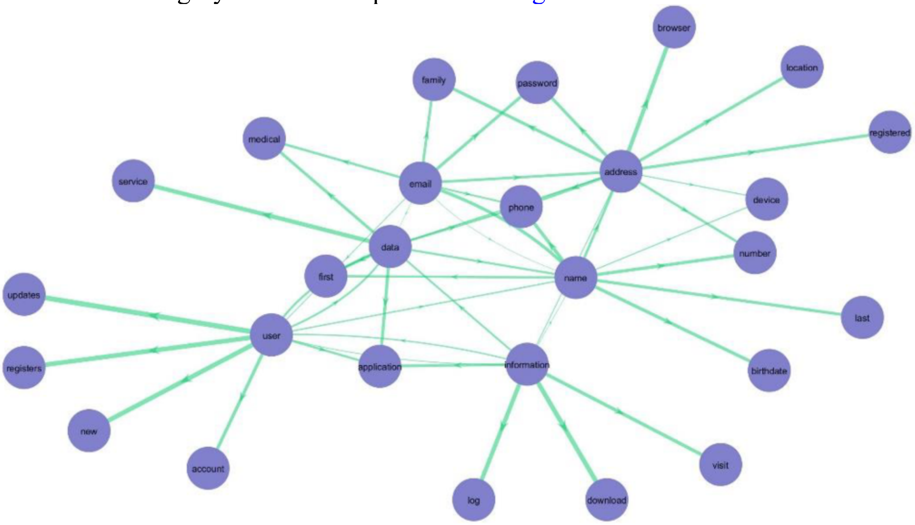


Figure 7: Visualization of co-occurrence analysis of parts of privacy policies – category "medical"

The visualization above shows that apps, not surprisingly, collect information about users, such as names, addresses, email, and alike.

However, a detailed inspection unveiled some interesting words worth further investigation. These were "location," "service," "device," "browser," "password," "medical," "family," and "doctor."

We separately visualized co-occurrences of the above words in two groups, the first four and the last four words together, see Figure 8 and Figure 9 and, respectively.

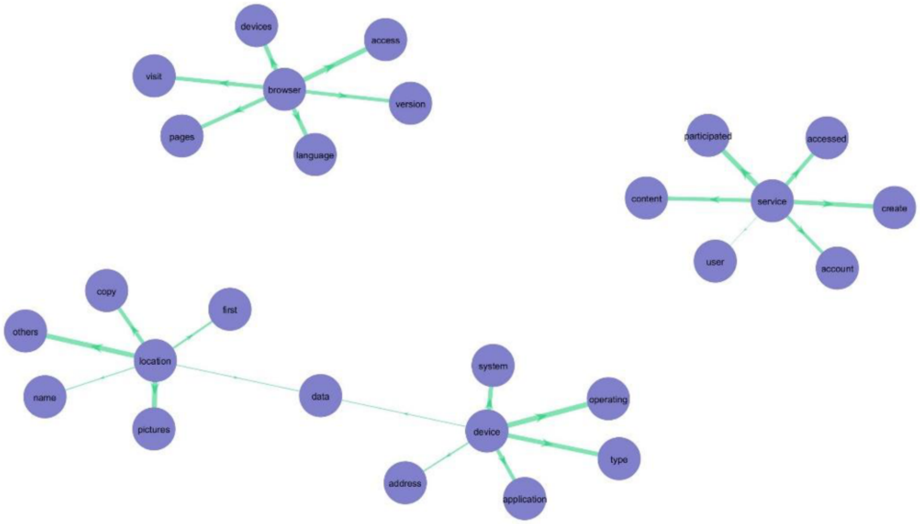


Figure 8: Visualization of co-occurrence analysis, words "location," "service," "device," and "browser." (medical apps)

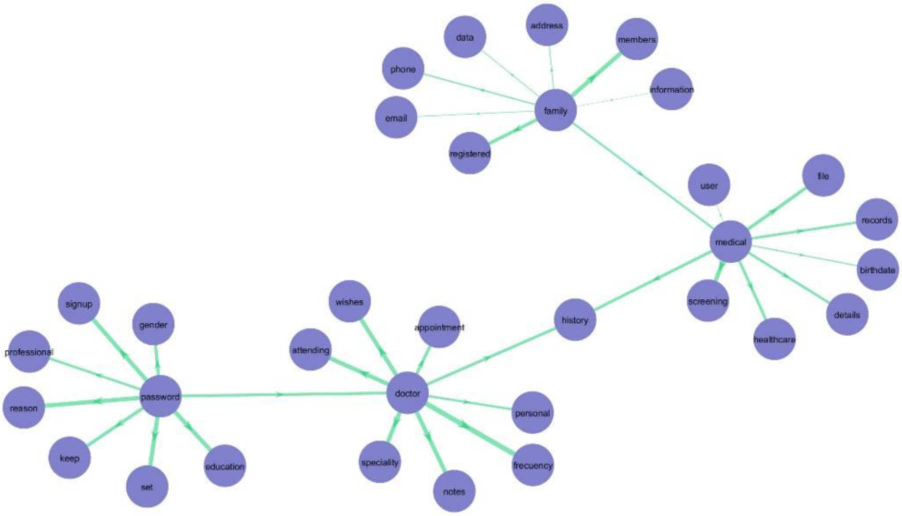


Figure 9: Visualization of co-occurrence analysis, words "password," "medical," "family," and "doctor." (medical apps)

Here, worrying patterns have shown up. Firstly, it is evident that apps store browser-related data, such as page visits, versions, language used, and devices being used. About devices, they collect their make and type, operating system used, and address (geolocation). Secondly, apps track the services being used and their content.

Apps collect data about the users and their family members, such as their emails and phone numbers, further enabling linking data from various sources.

Additionally, medical data are being collected, such as medical files, records, details, and screenings. Closely related are the data about doctor visits and appointments, together with visit frequencies and doctor's notes taken.

3.3.2. Co-occurrence analysis of the "Health & fitness" privacy policies

The visualization result of the co-occurrence analysis of the keywords of privacy policies in the category "health & fitness" is presented in Figure 10.

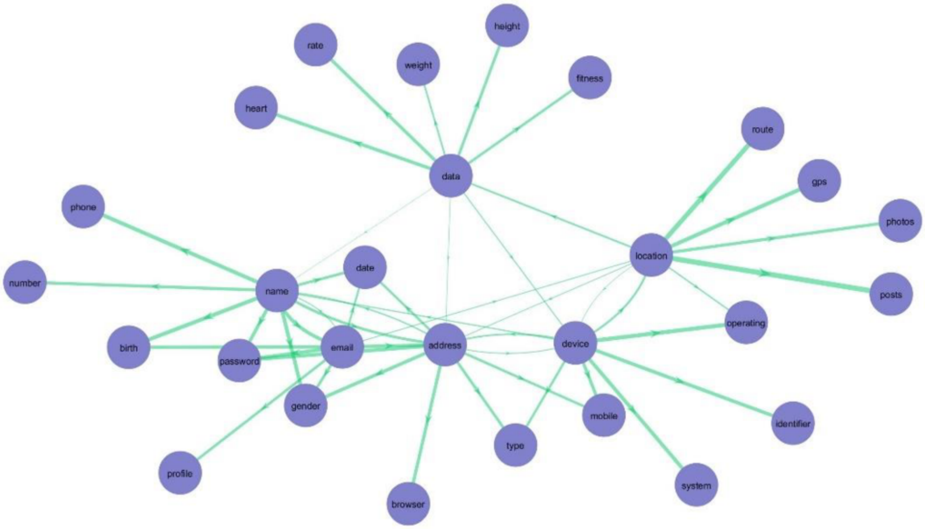


Figure 10: Visualization of co-occurrence analysis of parts of privacy policies – category "health & fitness"

The visualization above shows that apps, just as in the "medical" category, collect general user information (names, addresses, email, phone number). Same as before, more detailed information about users regarding "location," "service," "device," "browser," "password" pop up. Here, we do not repeat the co-occurrence analysis for the first group of keywords related to location and device/browser; the results are comparable to the medical group.

However, in the "health & fitness" group, we found the following keywords worth further investigation: "route," "heart," "weight," (and "height"), "gender," and "photos." From the word cloud, we add keywords "body," "activity," and "exercise" for further investigation. We separately visualized co-occurrences of the above two groups of the words.

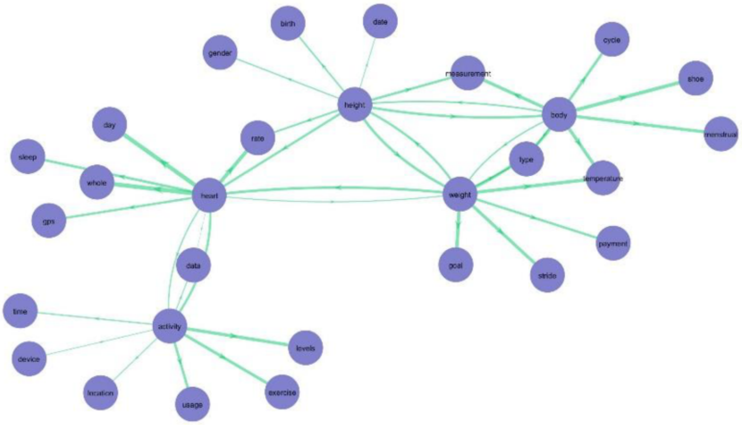


Figure 11: Visualization of co-occurrence analysis, words "activity," "heart," "height," "weight," and "body." (health & fitness apps)

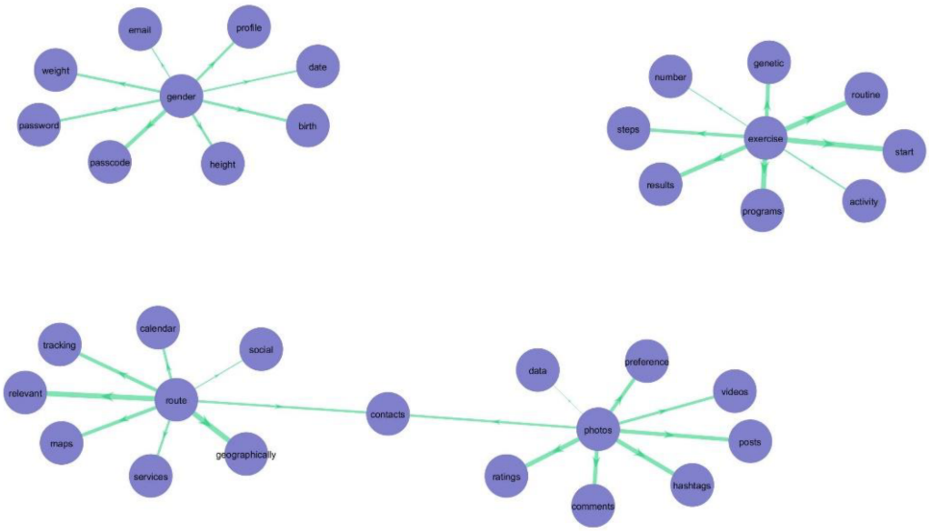


Figure 12: Visualization of co-occurrence analysis, words "route," "gender," "exercise," and "photos." (health & fitness apps)

Figure 11 and Figure 12 show a more detailed co-occurrence analysis of the selected words.

Here, interesting patterns show up. Namely, apps track many aspects of users' (sports) activities, including time, geographical location, and duration. Additionally, they track sleeping patterns and heart rates. Worryingly, they keep track of menstrual cycles and body temperatures. As expected, apps store photos, videos, comments, ratings, and preferences.

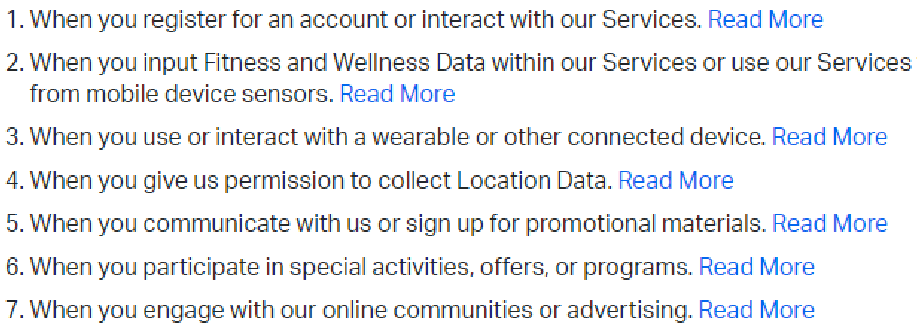
Both groups commonly store passwords, which poses a considerable security risk [26] and is directly against good password management policies [27]. It is a known fact that users reuse their passwords across several platforms and services. In the RTB ecosystem, where companies share sensitive data, including users' passwords, such a practice opens a whole new perspective on how securely data are stored and how easy it actually is to gain unauthorized access to users' data on other platforms.

4. Limitations

The first and most obvious limitation is the data source, being only the Google Play store. However, based on Statista's surveys [28], the Google Play store has more than 50 % market share, the rest of the market being shared by the Apple App Store, Windows Store, and Amazon Appstore. Additionally, we limited our research to the top 30 applications by the number of downloads from each group, "medical" and "fitness & health."

We believe that apps that are most frequently downloaded exist in both major mobile app stores. However, this is an assumption, and we have not checked it.

The following limitation is the automatic downloading and manual processing of privacy policies. For example, automatic downloading could have retrieved only a limited portion of the privacy policy, as per example in Figure 13:

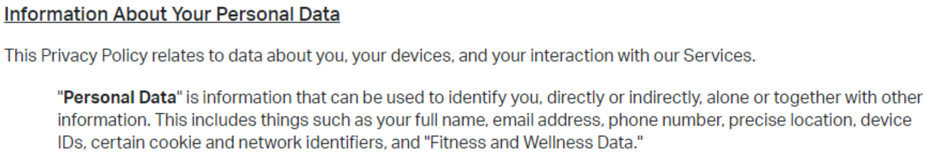


1. When you register for an account or interact with our Services. Read More
2. When you input Fitness and Wellness Data within our Services or use our Services from mobile device sensors. Read More
3. When you use or interact with a wearable or other connected device. Read More
4. When you give us permission to collect Location Data. Read More
5. When you communicate with us or sign up for promotional materials. Read More
6. When you participate in special activities, offers, or programs. Read More
7. When you engage with our online communities or advertising. Read More

Figure 13: Partial privacy policy

As seen from the example above, one has to click on the link "Read more" to see (and retrieve) the whole text.

The next limitation was the way we manually processed the privacy policies. Each one was searched for the section explaining what data the company is collecting – see Figure 14.



Information About Your Personal Data

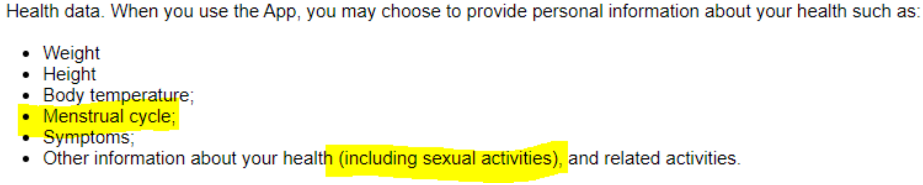
This Privacy Policy relates to data about you, your devices, and your interaction with our Services.

"**Personal Data**" is information that can be used to identify you, directly or indirectly, alone or together with other information. This includes things such as your full name, email address, phone number, precise location, device IDs, certain cookie and network identifiers, and "Fitness and Wellness Data."

Figure 14: A section of a sample privacy policy describes what data company is collecting

However, we noticed that companies are "hiding" relevant data to other sections of privacy policies and only list "by example" what they are collecting. None of the companies disclosed a complete list of attributes they are collecting; mainly, they explained the private data they gather by the most apparent attributes (name, email, location, IP address, etc.).

Finally, this work was intended to show the general trend and find attributes of personal information that are most frequently used. For this reason, our study does not reveal rare practices, bad practices, or outline particular private data some companies are collecting. One such example is a company offering a loose weight app for men – weight loss in 30 days, collecting the following data (Figure 15).



Health data. When you use the App, you may choose to provide personal information about your health such as:

- Weight
- Height
- Body temperature;
- Menstrual cycle;
- Symptoms;
- Other information about your health (including sexual activities), and related activities.

Figure 15: An excerpt from the privacy policy of an app for men collecting very specific data

Finding out what "other information" really means remains mission impossible.

5. Discussion and conclusion

This paper analyzed the texts of a sample of 62 privacy policies of the most frequently used free medical apps in the Google Play store, focusing on privacy policies describing which personal data are being collected and processed by the apps.

We chose free apps due to the prevailing business model in apps markets where developers make their apps available for free and get paid by advertisements shown inside the apps. A whole ecosystem was built around the advertisements on the web and inside apps, making it a multi-billion-dollar business, predominantly controlled by a single company.

The advertisements are served through a real-time bidding (RTB) process where a multitude of companies are collecting, storing, sharing, and doing business with private personal data. Due to regulations and restrictions in place, the burden of rules compliance is put on the apps' developers, forcing the users to accept the privacy policies, knowing that users neither read nor understand them. The users actually do not know what type of personal data the applications are collecting and further (for the purpose of advertising) sharing with a myriad of companies.

The paper's novelty is applying an automated linguistic analysis on privacy policies coupled with visualization techniques to unearth (deliberately!?) hidden information on which types of private data apps are collecting. To the best of our knowledge, no previous research on privacy policies was done using the specific linguistic technique combined with visualization.

Based on the co-occurrence analysis of privacy policies, a form of text mining, we have first visualized the results, followed by a manual inspection of simple graphs. Generally, medical, and health & fitness apps collect all the "expected" private data, such as user / personal names, gender, email, and physical addresses. Additionally, they collect and store photos, videos, chats, messages, opinions, and alike, all created by users within the apps offering such options.

However, a detailed analysis has revealed that apps also collect and store passwords (which is a hazardous practice since users reuse passwords across many websites, services, and apps).

Additionally, medical apps collect and process sensitive medical details, such as medical records, files, notes, doctor's appointments with their frequencies, and alike. It is incredibly worrying that apps collect data about the app users' family members and their medical data.

Health and fitness apps are, in general, and at the overview level, less privacy-invasive. They also collect health-related data, such as sleeping patterns, menstrual cycles, heart rates, weight and height, and bodily temperatures.

The future work can take several strains. Firstly, because data from individual users from different apps and their browsing activities are being shared in the background, thus creating a rich database, we could try to create a consolidated list of all possible attributes potentially being shared about an individual user. Secondly, we could extend our analyses not only to the co-occurrence of words, i.e., two words occurring at the same time, but also longer sequences and combine them into a single picture.

Sadly, lenient regulations are not protecting users in the app markets (neither in the browsing). A simple remedy would require putting personal data on a forbidden list, making trafficking with personal data a criminal offense. It cannot be expected that users first read what is being collected and then opt out of advertising in every single app they are using. Big players in the market have made the process of protecting personal data deliberately tricky, if not impossible. The care (and the burden) of privacy protection is put on users, which "By using the Service, you agree to the collection and use of information in accordance with this Privacy Policy." Nobody really reads what's collected. And stored. And shared.

A simple by default rule, "no personal data storing & processing," could be a good start.

Acknowledgment

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-0057) and the University of Maribor (www.um.si, core funding).

References

1. Brumen, B., et al., *Use of Mobile Technologies in Tourism: Natural Health Resorts Study.* Mediterranean Journal of Social Sciences, 2020. **11**(4): p. 1.
2. Clement, J. *Mobile internet traffic as percentage of total web traffic in October 2020, by region*. 2020 2021-02-01]; Available from: https://www.statista.com/statistics/306528/share-of-mobile-internet-traffic-in-global-regions/.
3. Apple. *Choosing a Business Model*. 2021; Available from: https://developer.apple.com/app-store/business-models/.
4. Gui, J., et al. *Truth in Advertising: The Hidden Cost of Mobile Ads for Software Developers*. in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. 2015.
5. Ruiz, I.J.M., et al., *Impact of Ad Libraries on Ratings of Android Mobile Apps.* IEEE Software, 2014. **31**(6): p. 86-92.
6. Yuan, S., J. Wang, and X. Zhao, *Real-time bidding for online advertising: measurement and analysis*, in *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*. 2013, Association for Computing Machinery: Chicago, Illinois. p. Article 3.
7. Brumen, B., *Automated Text Similarities Approach: GDPR and Privacy by Design Principles*, in *Information Modelling and Knowledge Bases XXXII*, M. Tropmann-Frick, et al., Editors. 2021, IOS Press. p. 213-226.
8. Brumen, B., et al., *Outsourcing medical data analyses: can technology overcome legal, privacy, and confidentiality issues?* J Med Internet Res, 2013. **15**(12): p. e283.
9. Allen, A.L., *Privacy Law and Society.* 1st ed. American Casebook Series. 2007: Thomson West.
10. Allen, A.L., *Privacy and Medicine*, in *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, E.N. Zalta, Editor. 2011, Stanford University: Stanford, CA, USA.
11. Solove, D.J., *A Taxonomy of Privacy.* U Pa L Rev, 2006. **154**(3): p. 477-564.
12. EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.* Official Journal of the European Union, 2016. **L:2016:119**.
13. Chabinsky, S. and P.F. Pittman, *USA*, in *The International Comparative Legal Guide to: Data Protection 2019, 6th Edition*, N. Catlin, T. Hickman, and D. Gabel, Editors. 2019, Global Legal Group: London, UK.
14. Google. *Helping publishers comply with the California Consumer Privacy Act (CCPA)*. 2021 2021-02-03]; Available from: https://support.google.com/adsense/answer/9560818?hl=en.
15. Kaldestad, Ø.H. *250,000 words of app terms and conditions*. 2016; Available from: https://www.forbrukerradet.no/side/250000-words-of-app-terms-and-conditions/.
16. Phillips, A.M., *Research Handbook on Law and Courts*, in *All your data will be held against you: secondary use of data from personal genomics and wearable tech*. 2019, Edward Elgar Publishing.
17. Cyphers, B. *Google Says It Doesn't 'Sell' Your Data. Here's How the Company Shares, Monetizes, and Exploits It.* 2020 Electronic Forntier Foundation, March 19, 2020]; Available

from: https://www.eff.org/deeplinks/2020/03/google-says-it-doesnt-sell-your-data-heres-how-company-shares-monetizes-and.

18. Zhang, W., et al., *Real-time bidding benchmarking with ipinyou dataset.* arXiv preprint arXiv:1407.7073, 2014.

19. Google. *Introduction to Open Bidding*. 2021 2021-02-05]; Available from: https://support.google.com/admanager/answer/7128453?hl=en.

20. Cyphers, B. *Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance*. 2019 Electronic Forntier Foundation, December 2, 2019]; Available from: https://www.eff.org/wp/behind-the-one-way-mirror#Real-time-bidding.

21. Google. *Helping publishers comply with the California Consumer Privacy Act (CCPA)*. 2021 2021-02-05]; Available from: https://support.google.com/adsense/answer/9560818?hl=en.

22. Gledhill, C.J., *Collocations in science writing*. Vol. 22. 2000: Gunter Narr Verlag.

23. Wartena, C., R. Brussee, and W. Slakhorst. *Keyword extraction using word co-occurrence*. in *2010 Workshops on Database and Expert Systems Applications*. 2010. IEEE.

24. Matsuo, Y. and M. Ishizuka, *Keyword extraction from a single document using word co-occurrence statistical information.* International Journal on Artificial Intelligence Tools, 2004. **13**(01): p. 157-169.

25. MathWorks. *Co-occurrence analysis with Text Analysis Toolbox*. 2020 [cited 2020-02-05; Available from: https://github.com/mathworks/Co-occurrenceAnalysis-and-visualization.

26. Brumen, B., *System-Assigned Passwords: The Disadvantages of the Strict Password Management Policies.* Informatica, 2020. **31**(3): p. 459-479.

27. Grassi, P.A., et al., *NIST Special Publication 800-63B. Digital Identity Guidelines. Authentication and Lifecycle Management*. 2017, National Institute of Standards and Technology: Gaithersburg, MD, USA.