Automated Extraction of Structured Data from the Social Network Instagram

Petr Františ, Michal Bureš, Aneta Coufalíková and Ivo Klaban
Department of Informatics and Cyber Operations, Faculty of Military Technology, University of Defence, Brno, The Czech republic

Petr.Frantis@unob.cz
Michal.Bures3@unob.cz
Aneta.Coufalikova@unob.cz
Ivo.Klaban@unob.cz

Abstract: The paper explores the extraction of structured information from the social network Instagram through a suitable application programming interface, namely the unofficial Instagram Private API. It focuses on creating a computer program that identifies which posts a user has tagged as "Likes" and then stores this information for profiling specific user profiles. The introduction of the paper highlights the general use of social media in modern society and the importance of personal data for these platforms. It specifies the aim of the study, which is to extract information from Instagram and then analyse it for user profiling. It then describes the evolution of the social network Instagram and key features such as different types of posts. This paper further focuses on the solution and implementation by using Python programming language to minimize the load on Instagram servers and reduce the risk of detection of automated processes. It describes the process of setting up new Instagram accounts, the obstacles in obtaining login credentials, and the need to simulate human behaviour to bypass the network's defence mechanisms. It then focuses on the actual retrieval of information such as the users followed, their posts and information about which posts the user has marked as favourites. It mentions that extracting data from closed profiles is difficult and elaborates on the technical challenges associated with this task. A significant part of this paper is a discussion of Instagram's defence mechanisms that respond to automated computer programs. It describes access denial, account blocking, and identity verification prompts such as CAPTCHA tests. Finally, the conclusion summarizes the results obtained, which indicate the acquisition of approximately 90,000 records for user profiling. It discusses the shortcomings of a fully automated solution due to Instagram's account creation conditions and defence mechanisms. It mentions the need for further research and highlights key gaps and challenges in this area. Overall, the study highlights the technical and security challenges in extracting information from Instagram and emphasises the need for further research and improvements in the technical procedures for extracting data from the platform.

Keywords: Instagram, Profiling, Instagram Private API, Automation, Osintgram, Python

1. Introduction

Social networking is a prevalent phenomenon in contemporary society that almost every individual utilises (Eurostat, 2021). While some networks are limited to a single country, others serve as transnational platforms, connecting diverse cultures and schools of thought in one place. However, a fundamental aspect of all social networks is that users systematically contribute significant amounts of data every day (Marr, 2018).

Social network users often share highly sensitive information, such as their whereabouts, favourite leisure activities, lists of friends, and other personal details. This information is processed and analysed by each social network, which then uses it to implement new functionalities and to influence the behaviour of its users, either directly or indirectly. Due to its significance as a major market value for social network providers, structured public access to this data is not available.

This paper aims to extract structured data from Instagram, a social network, with a focus on individual user information. Each user generates activity on this network. By assuming that they identify with the content they label as 'Likes', it is possible to create a psychological profile of the individual. This study focuses on developing a computer program that extracts this information from Instagram and makes it available for detailed analysis to profile a specific user.

2. Instagram

The social network was launched in 2010 as a mobile app for the iPhone 4, available in the App Store. In 2012, Instagram released a version for the Android operating system, leading to a rapid increase in its user base. Shortly after its acquisition by Facebook, Inc. (now known as Meta Platforms, Inc.) for approximately 20 billion Czech crowns, Instagram underwent significant changes that greatly influenced its future development (Eldridge, 2024).

This change in ownership between two software-focused companies resulted in significant changes to the approaches and features offered to users. A separate chapter focuses on the application programming interfaces that underpin these approaches. Users are attracted to this social network by key features such as image sharing, post creation and monitoring, and interactive features like comments and likes (Meta, 2024).

Posts are the primary means of communication for each user and are currently divided into five categories: Image, Video, Carousel, IGTV Video, and Sequence (see Table 1).

Table 1: Categories of contributions

Name	Description
Image	Photo most often taken with a mobile phone.
Video	Video most often recorded with a mobile phone
Carousel	Containing images and videos displayed sequentially after a certain time interval
IGTV Video	Longer-form video shot vertically with a mobile phone for up to 60 minutes
Sequence	Fun short video up to 15 seconds in length

One important feature of a post is its ability to link to another user. The owner can tag one or more other users, even if they have no connection to the post. Another way to associate a user with a post is to use the hashtag symbol in the caption or to collaborate with them using a double post. With a double post, you can upload a post to both profiles simultaneously. In an Instagram post, the owner can tag a user with whom they are not friends. Once the post is published, the tagged user will receive a notification of the tag.

In social media, users can write captions for their posts and include hashtags to link them with other posts. Posts may also include location information and allow for comments from other users.

2.1 Stories

For a period of 24 hours, any user has the ability to upload a story, which can be an image, video, or link to another post. These stories can be accessed by clicking on the user's profile picture and are automatically saved to the archive. After the initial 24-hour period, users have the option to keep their stories visible by featuring them as highlighted stories on their profile.

These stories can be shared or forwarded if the author has enabled this option. They can also be marked with a heart symbol to indicate approval of the story. However, this information is not public and is only available to the author of the post. By default, the poster receives notifications from the mobile app when a user marks their story with a heart symbol, indicating sympathy.

2.2 Watchers and Watched

The platform also includes features that allow users to connect directly, but these interactions are tracked and monitored. The followers of a profile are users who have clicked the 'follow' button on a public profile or, for a private profile, those who have been granted permission to follow by the profile owner. Similarly, tracked users are those who have accepted a tracking request on a private profile or have a public profile and have been tracked by the user.

This feature enables additional functionality, such as displaying the followed user's story on the main page after logging in, receiving notifications when the followed user posts, and displaying the post on the main page.

The main source of interaction and connection between users in this work is through watching and tracking. We assume that a user follows those whose activity interests them and directly or indirectly influences their thinking. Social graphs can be used to monitor these interactions and look for links between the set of users under study.

2.3 Direct Messaging

Every social network involves sharing text information between two users. Instagram allows users to send links to posts, stories, and profiles, in addition to text and emojis. Users can also send photos, temporary photos, voice messages, or initiate a call or video call.

Regarding the aim of this paper, this feature is not significant because, for user privacy reasons, only those users directly involved in a conversation can access this information. At present, there is no way to identify which users are messaging each other and which are not.

3. Application Programming Interface

When extracting data from web servers, the standard procedure is to use queries. In this study, we aimed to identify and evaluate the available application programming interfaces (APIs) that could facilitate the data extraction process and enable more efficient system integration. This approach aligns with current trends in software engineering, where system integration plays a crucial role in tool innovation (Jensen et al, 2021).

The aim of our study was to identify user data related to social media interactions, specifically data on posts that users marked as 'Likes'. This information is significant for user profiling, which is important for sociological research and marketing analysis. Our research found that certain APIs provide access to this specific data, which was crucial for our purposes. The following sections present detailed analysis and comparisons of several interfaces that were considered.

3.1 Instagram Graph API

The official application programming interface is exclusively designed for special user accounts designated as business or creative upon request (Meta, 2024). Its main functions include uploading posts, replying to comments, filtering hashtags, and obtaining statistical data on profile status. Processing information about other accounts not owned is not enabled, making this interface unsuitable for achieving the goal of this thesis.

3.2 Instagram Basic Display API

An additional option for acquiring official data is the basic imaging interface (Meta, 2024). This interface can retrieve the same information as the previously mentioned interface but is designed for common profiles that are not marked as business or creative. However, it is limited to retrieving information about the logged-in account and cannot access information about other accounts, making it unsuitable for our purposes.

3.3 Other Official Interfaces

The creators of the social network Instagram provide other public interfaces, such as interfaces for direct messaging, sharing stories or posts, and using links to other web applications (Meta, 2024). However, none of these interfaces met the requirements of our work.

3.4 Osintgram

The tool under analysis is named by combining 'OSINT', an acronym for 'open source intelligence', with a suffix derived from the name of the social network Instagram (Datalux, 2022). It was created by a user with the nickname Datalux on the GitHub platform and is capable of extracting useful information from Instagram using the command line. GitHub is a widely used platform for version control and sharing software projects in the developer community.

The tool's interface allows users to retrieve information about other Instagram users but requires authentication through an existing account. The documentation warns that using the tool may trigger Instagram's security mechanisms, which require authentication to confirm that interactions are by a human and not automated software.

The tool offers various options such as a list of followers, emails, posts, and user stories. However, it does not provide information about users who have tagged a particular post as 'Like'. For our study, this functionality proved to be insufficient.

3.5 Instagram Private API

The interface is once again available on GitHub and is a crucial component of the Osintgram source code. This interface is highly functional and well-documented, making it effective for research and analytics applications (Ping, 2019). However, there is a risk of suspicious activity detection when interacting with Instagram servers, which may result in rejected queries and loss of access to the associated account.

The study primarily chose the Instagram Private API due to its web interface that can be accessed without logging in. The source files are written in Python and use GraphQL, a query language used by Instagram servers. Additionally, the API allows logging in using stored cookies, reducing the likelihood of detection by the server's defence mechanisms. These cookies usually have a validity of two months.

Other benefits of the interface include its ease of use. Once a name and password are entered, the interface authenticates with Instagram's servers, providing access to a variety of features. However, to obtain specific information, such as lists of users who have tagged posts as 'Likes', a universally unique identifier (UUID) must be entered, which is obtained with the login cookies.

However, the interface is unofficial and may be prone to malfunctioning if Instagram's query structure changes. Furthermore, Instagram may flag access through this interface as a security risk due to the use of Android devices. This is because the geographical location of logins can vary, which may indicate the use of different Android devices over time. It is important to note that this risk is particularly high because of the dynamic change of the login configuration file.

4. Solution and Implementation

Python was chosen as the programming language for the presented software solution due to its compatibility with the selected API. The software design aimed to minimize queries to the Instagram server to reduce its load and the risk of detection by its defence mechanisms. This approach included introducing artificial delays between queries, ranging from 20 to 120 seconds. This helped to increase the time required to retrieve data while also reducing the likelihood of detection.

Additionally, emphasis was placed on efficiently processing and storing the large volume of data collected from the server. To avoid automatic detection by Instagram's defence mechanisms, authentic human behaviour was simulated. It was suggested that an alternative and potentially more effective approach could be to include queries mimicking human clicks in the web browser. However, due to the high complexity of defining and replicating average human behaviour, this approach was not implemented in our research.

The study aimed to create a computer program that could identify Instagram posts tagged as 'Likes' based on the user's username. This was accomplished by gathering and analysing data on the monitored users, their posts, and the users who tagged those posts. Then, an analysis was conducted to identify the specific username entered in the input. The methodology and implementation steps of this solution are described in detail in the following sections of the paper.

4.1 Setting up an Instagram Account

To conduct the research, it was necessary to register an account on the social network Instagram. This required providing certain information, including a unique username, password, full name, and either a mobile phone number or email address. As the study focused on the creation of a botnet, the potential for automated generation of this data was investigated. The main challenge was verifying an email address or phone number, which Instagram requires through a six-digit verification code.

The experiment tested disposable internet phone numbers with an accessible SMS mailbox that are freely available online. Although Instagram confirmed sending the verification code, the message was not found in the incoming messages. The evidence indicates that Instagram may need additional verification from the operator to confirm legitimate use of the phone number. It is possible to have up to five user accounts per phone number, according to available sources.

Disposable email boxes were also investigated, but they faced the same problem of unavailability of the authentication code. This phenomenon may be due to the need to verify the email address with the domain owner or an Instagram-owned database of suspicious emails. Current email security trends require two-factor

authentication, typically via a phone number. Therefore, the most effective way to set up new Instagram accounts is to use a phone number. In our case, we implemented this option by purchasing a new SIM card.

To retrieve information through a new account, it was crucial to ensure that it was indistinguishable from regular user accounts. Therefore, we developed an identity creation strategy that involved tracking designated profiles and interactions with posts. However, Instagram restricted the activity of our accounts after tracking an average of seven profiles, often temporarily blocking them for a week. This phenomenon highlights the advanced mechanisms used to detect automated behaviour on this platform.

It is worth noting that shortly after setting up our accounts, one of them started monitoring a suspicious account, which suggests the possible existence of a computer program on the other end. The account was flagged as suspicious based on the content and timing of its interactions, providing evidence for the existence of sophisticated automated accounts on this platform.

4.2 Finding the Monitored

The first step of the automated computer program was to identify the users tracked by the target Instagram account. To achieve this, we used the Instagram Private API, which provided usernames and other information in individual server responses. As each response contained a large amount of data, we maximized its use for subsequent profiling and stored it in the SQLite3 database.

The Instagram Private API utilises a sequential data loading mechanism to reduce the amount of data in each request. As a result, a single response can process approximately 100 profiles when retrieving information about tracked users. The following data was stored in the database from these profiles:

- unique user identification number (pk_id),
- username,
- full name,
- information about whether the profile is private or public,
- URL of the profile picture.

Such publicly available data can be collected and processed, but it is important to respect data protection laws and only use it for legitimate purposes. These laws may vary depending on the location.

It should be noted that our automated program was unable to retrieve information such as posts from private accounts. The limited fact prevented the extraction of information regarding the target user's interactions, including the 'Like' label of a post. It is challenging to bypass Instagram's privacy protection mechanisms in the current technological and legal environment, and doing so could potentially violate privacy rights.

Regarding the persistence of stored data, it is important to note that the URL of profile images is temporary. Storing the images themselves in a database may be more efficient but would significantly increase data volume.

4.3 Extracting Data from the Posts

The second step in profiling a specific Instagram user was to gather data from the posts of users followed by the target account. This data was collected using the Instagram Private API, which provided both the necessary data and additional information that was not the primary focus of the study. The database stores specific attributes extracted from the data, such as post type, unique post code, number of comments, number of 'Like' tags, location name, address, city, latitude and longitude, post URL, and caption. Storing the posts themselves in the database was considered, but due to the temporary nature of the URL, it would have significantly increased the data volume.

Due to the large amount of data contained in server responses, they were often divided into multiple parts. When retrieving contribution data, approximately 20 contributions could be processed in a single request. In some cases, submissions did not include complete information, particularly regarding location if the author chose to remove this metadata. In these instances, NULL values were used to replace the missing information in the database. Significant problems with duplicate contributions were noted, particularly when one of the authors was a private user, making it impossible to retrieve additional information. As a result, the number of contributions retrieved was not always consistent with the expected number from the queries.

When copying information from a social network, it is important to consider the frequency of database updates and the rate at which information changes. In this version of the program, the database was not updated

continuously due to complications associated with detecting and processing hidden double posts from private users.

4.4 "Likes" Posts

The last step of the profiling process involved gathering data on users who labelled particular Instagram posts as 'Likes'. This information was then filtered and analysed using the methodology outlined in section 4.2. Throughout the study, the Instagram terms of use were updated to permit users to conceal both the number of 'Likes' and the list of users who tagged the post. The change resulted in Instagram displaying only the first 100 users who had tagged the post.

To address this, we applied the user tracking feature in the research tool. If a particular user was being tracked by our profile, they were displayed among the first 100, ensuring that relevant data was captured. The captured information was stored in the database using two keys: the username and the unique post code. The database was efficiently filtered by a specific username, allowing for user profiling and meeting the research objectives. Around 90,000 records were retrieved during the study. Figure 1 illustrates an example of the filter applied to the username 'david'.

	username	code
	david	Filter
1	krajnak_david	Ckth5TzNBkE
2	krajnak_david	CkZSP47Nl4B
3	krajnak_david	CkZQTX0N_N5
4	krajnak_david	CkZPtC4tIEL
5	davidsojka20	B5lYMfjFWap
6	davidberger93	BfRFixUHP0N

Figure 1: Database table contains information about which user has marked a post as 'Like'.

Analytical graphs of the selected person's interests can be automatically generated based on the frequency of keywords (hashtags) used in their posts (see Figure 2).

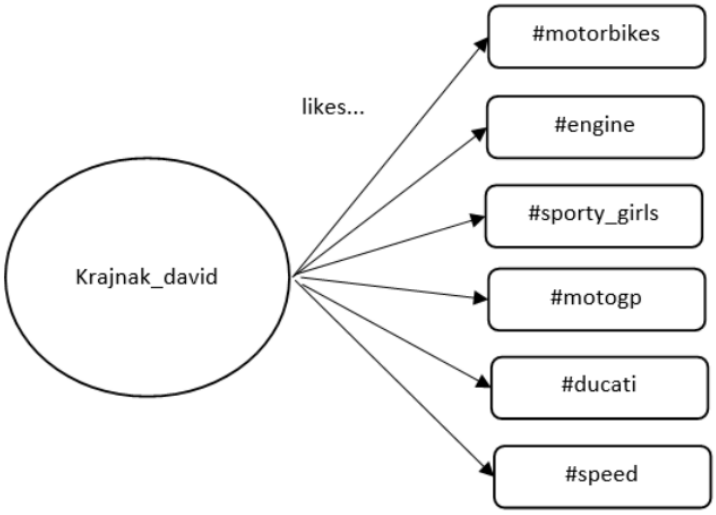

Figure 2: Graph of the selected person's areas of interest

The data from the generated database can be used to perform more sophisticated analyses of individual relationships based on the number of 'like' posts tagged to each other. These relationships can then be displayed in the form of mind maps, either static or interactive with appropriate tools, allowing the user to navigate through the map of individual relationships (see Figure 3) (Gould et al, 1986).

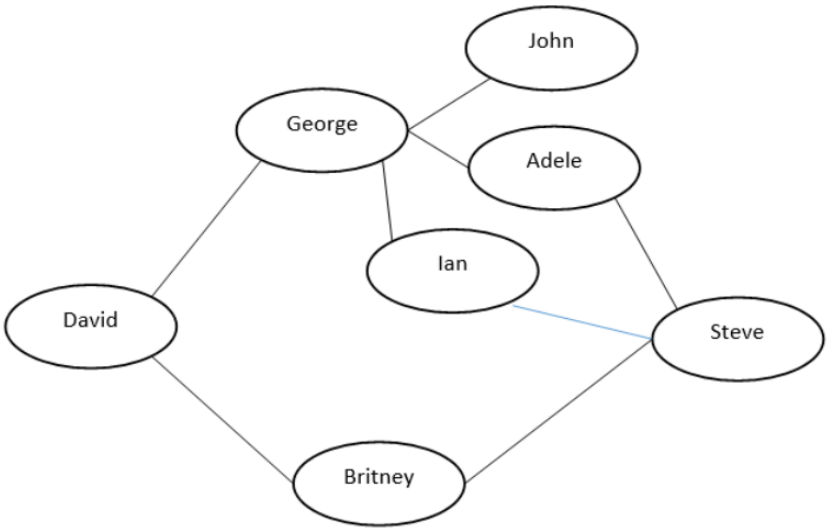

Figure 3: Relationship mind map (anonymised)

5. Instagram's Defence Mechanisms

Similar to other social networks, Instagram has implemented various defence mechanisms to prevent automated computer programs. These mechanisms are intended to safeguard the server from excessive computing power through queries, prevent the publication of inappropriate posts, prevent users from spamming, and prevent other activities that violate the platform's terms of service.

Throughout our research, the computer program we used was frequently rejected and logged out due to suspicious behaviour or excessive server usage. For instance, a server response with code 401 indicated a user logout, while code 429 indicated a need to slow down the frequency of queries. The maximum number of queries allowed varied depending on the rejection history of a particular account, usually around 20 queries. It appears that Instagram monitors and analyses the behaviour of accounts that are considered suspicious and adjusts their access policy accordingly.

To partially solve the issue of repeated rejections, a new login session can be generated, as the timeout limitation seems to be associated with the universal identifier of the session. However, it is important to note that multiple login sessions linked to a single user account may be perceived as an indication of automated behaviour.

One response to the query was the display of a prompt, which is a security measure that requires human confirmation. In this case, a CAPTCHA test was presented, which required the selection of specific images based on a text input. The computer account was blocked multiple times during the data retrieval process due to suspected automated behaviour. To unblock the account, specific requirements must be followed:

- Write the username and specific code on a blank piece of paper.
- Take a photo holding this paper with your face and hand visible.
- Upload this photo to the form and wait for the evaluation.

None of our research profiles had a profile photograph, and we were unable to provide a face or hand for security reasons. Therefore, we were unable to circumvent this process.

6. Conclusion

The study aimed to develop and test a computer program that extracts information from Instagram using the unofficial Instagram Private API. Specifically, the program collected data on posts that users had tagged as 'Likes'. The testing phase successfully stored approximately 90,000 entries in the database.

The collected data provides valuable information for subsequent user profiling, which has potential applications in areas such as marketing, social research, or trend analysis. During the program's development, we identified and analysed various obstacles that limit the possibility of creating a fully automated solution. The obstacles encountered during the project primarily involved challenges related to creating new accounts and Instagram's defence mechanisms for detecting and blocking suspicious activity.

The program was developed with scalability in mind, allowing it to run continuously on multiple computers, resulting in increased efficiency and faster information retrieval. During the development process, we identified and documented certain performance shortcomings in the source code documentation. These deficiencies provide valuable input for further development and improvement of the program, allowing for even greater efficiency and reliability in the future.

This study makes a significant contribution to the field of social network data extraction and offers a useful basis for further research and development in this rapidly developing area. The findings and knowledge gained from this study offer valuable insights for other developers and researchers working on comparable projects in the realm of social media.

References

Datalux (2022) "Osintgram", [online], https://github.com/Datalux/Osintgram#osintgram-

Eldridge, A. (2024) "Instagram social networking service", [online], https://www.britannica.com/topic/Instagram

Eurostat (2021) "ICT use in households and among individuals - 2021: Individuals in EU countries using social networks evolution over time", [online], https://www.czso.cz/documents/10180/142872020/ 062004210705.pdf/4dc708a2-442e-41b8-bd43-e12a7ecc42c4?version=1.1.

Gould, P. and White, R. (1986) *Mental Maps*, 2nd ed., Routledge, London.

Jensen, M. and Patel, H. (2021) "Trends in Software Engineering: System Integration and Its Challenges", *Journal of Advanced Computing*, Vol. 12, No. 3, pp 234-247.

Marr, B. (2018) "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read", [online], https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Meta (2024) "Instagram", [online], https://about.meta.com/technologies/instagram/

Meta (2024) "Instagram Platform", [online], https://developers.facebook.com/docs/instagram

Ping (2019) "Instagram Private API", [online], https://github.com/ping/instagram_private_api#instagram-private-api