Determination of movement profiles based on open-source data from social media

Philipp Rosenkranz

Hamburg Police, philipp.rosenkranz@polizei.hamburg.de

Wilfried Honekamp

IACS, University of Applied Sciences Stralsund, wilfried.honekamp@hochschule-stralsund.de

Abstract:

Some crime cases cannot be investigated due to the discrepancy between their socially harmful nature and the level of intervention required to investigate them. Investigative approaches in connection with special collection methods of open-source data from social media could provide information that can usually only be obtained through measures with strong encroachments on fundamental rights. Social media offer an approach for law enforcement to quickly and directly access a large amount of openly accessible personal data of the suspect users.

This contribution focuses on the acquisition of open-source data from online users to determine movement profile. To this end, 90 digital identities were observed in their social media lives over 20 days. During the observation period, 4,186 contributions were recorded in five online services and then evaluated. A geotag was used for 834 entries and the location could be designated as known for 696 entries.

The open-source data from social media is suitable to determine movement profiles of the suspects. Much of the open-source data recorded in this investigation could be

used to create a movement profile for the corresponding digital identities. The results indicate that Instagram is currently best suited for creating movement profiles. Snapchat also offers great potential for creating movement profiles. User accounts can be easily assigned to users via Instagram in particular. While Facebook is only suitable to a limited extent, Twitter and YouTube are ruled out for the desired goal.

1 Introduction

Social networks and social media portals have become part of everyday life for many people these days. Their use seems to have become so self-evident that it is no longer questioned. Drivers look at their smartphones during red phases at traffic lights because they cannot seem to stand the time without information. Social media play a significant part in the heavy use of smartphones. Despite the ubiquitous presence of social media, their benefits do not seem to have found their way into all areas of professional life. An awareness of the power of social media in everyday professional life is also needed in order to profitably incorporate the benefits into one's job. This chapter will focus on the advantages and disadvantages of law enforcement in social media.

It is not far-fetched to think that investigators may also recognise further investigative leads that could very likely lead to the identification of the crime suspect or even to the clarification of the entire facts of the case, but must refrain from carrying out the criminal procedural measures required for this purpose, since these are not provided for by the legislature for all offense areas. If neither the police nor the public prosecutor's office are able to identify an offender using the classic measures available under criminal procedure, the public prosecutor's office will discontinue the criminal proceedings.

A readily available and much more resource-friendly way of obtaining information could be the evaluation of open-source intelligence (OSINT) from social networks. Investigative activities in social media could also be linked to significantly lower powers of intervention and thus be applied in a straightforward manner. In certain cases, there might not be any interference with the fundamental right of the respective user account holder at all, thus making an enabling basis obsolete.

2 Background

First of all, it is necessary to explain the concept of OSINT data, to address its evaluation and to explain which social media could be considered for criminal investigation on the Internet. Next, it is necessary to explain the legal requirements that the legislator has created.

"Publicly Available Information, which is data, facts, instructions, or other material published or broadcast for general public consumption; available on request to a member of the general public; lawfully seen or heard by any casual observer; or made available at a meeting open to the general public." (Morrow, 2006, p. 2-1)

Accordingly, OSINT data is any publicly available information that has been published by an originator for free disposal and where the originator no longer has control

over the continued use of that information. Accordingly, data from social media constitutes OSINT data.

According to the view taken here, Wehlte (2016, p. 118) has recognised that it is not so much the extraction of information from social media that poses a problem, but rather the evaluation and interpretation of the data obtained. The source is judged on the basis of the information it has provided so far or on how it is to be assessed based on the overall circumstances. For the evaluation of the information, it is not the personal assessment of the investigator that is decisive, but the proximity of the source to the information reproduced. This should enable a value-free assessment (Kroll and Schwarz, 2001, p. 215).

It is important to note that the communication channels in real and virtual groups differ significantly. In the real group, information is passed on through distributors before it reaches the user. In the virtual group, on the other hand, the information is passed on directly from the source to the user, and then discussed and passed on within the group. This is a complex and undirected form of information transmission (Labudde, 2018, pp. 10-11). The transmission of information does not only take place within a group. The transitions to other groups are fluid.

The increasingly intensive use of social media is leading to exponential data growth and an accompanying large data pool in social media. IT experts and market researchers refer to this as Big Data (Welker and Kloß, 2014, p. 37). The data volumes on the Internet fulfil three essential requirements for big data: an enormously large amount of data (volume), a variety of different information in the form of images, sound, video, and text (variety), and a high speed in the creation of new information, but also in terms of processing and analysis requirements (velocity) (Scheffler, 2014, p. 20). Big Data is actually intended to express that the volume of data cannot be analysed as a whole without further ado due to its size and/or other factors. Conversely, however, it also follows that one obtains numerous promising data for analysis if one only considers partial areas.

This finding is supported by several studies. According to these, social media are used by online users for more than an hour a day on average. The usage behaviour is more intensive on weekends than on weekdays. One in four online users operates their own channel on the social network and counts around 175 people among their interested users (followers/friends) (Tesche, 2018).

3   Methods

For this contribution, open-source data of 90 onliners was observed over 20 days in their social media lives. A methodology was developed for the empirical data collection and analysis, which is explained in the following five subsections. The following online services were selected:

- Instagram,
- Facebook,
- Twitter,
- Snapchat and
- YouTube.

In this contribution, the focus is on the collection of OSINT data from normal onliners for the purpose of motion profiling. In order to ensure the most detailed collection of all eligible digital identity data, a database was created for each digital identity. In each database, tables were created for the five online services. In these the following contributions were stored, that were published via an online service:

- date and time,
- story (yes/no),
- capture type,
- location (address),
- location (location in own words),
- activity,
- accompanying person and
- additional information.

It was assumed, that the determination of general activity could sometimes be difficult. If the digital identity meets with her family during her free time, the general activities of family and also free time could be considered. If everyone eats together, a third variation would be possible. The deciding factor in selecting the general activity was always what was in the foreground.

Examples:

1. The digital identity is on vacation at the beach with its family and is sunbathing.
2. The digital identity is eating pizza in a restaurant with its family on vacation.
3. The digital identity, who is not at home much, is sitting with her family on the couch in the living room playing a board game.

In the first example, the focus is on recreation on vacation away from everyday life. Accordingly, it is the general activity of "leisure". In the second example, the focus is on eating together on vacation. Here, the general activity "eating" is relevant. In

the third example, the focus is on spending time together with the family. For this reason, this should also be selected as a general "activity". The three activities can be described in more detail using the activity field. Thus, in the first example the activity would be described as "sunbathing," in the second example as "eating pizza," and in the third example as "playing board games".

Data collection took place from 11:19 a.m. on April 21, 2019, to 10:10 p.m. on May 10, 2019. During this time, all published posts of the previously mentioned digital identities were recorded in the respective available online services and entered into the corresponding databases. In doing so, care was taken to ensure that data collection was as seamless as possible, with only unavoidable delays caused by general human needs coming into play.

For the statistical analysis of the collected data, a profile was created for each digital identity in the form of a collection of tables including continuous text on the further findings about the digital identity, the recognised habits and other special features about the digital identity. In addition, a summary table collection was created to illustrate the statistical results of the five groups of digital identities.

4   Results

During the observation period, 4,186 entries were recorded in the five online services and subsequently evaluated. A geotag was found for 834 entries and the location was known for 696 entries. YouTube was not considered further because none of the 90 onliners was active there during the period under review. One user was active on Twitter, but his 7 entries referred exclusively to his new Instagram account.

4.1   Instagram

Instagram recorded the highest usage rate among all online services. This result confirms the study on social media usage behaviour. It also takes into account the fact that every digital identity has an Instagram user account, but not user accounts for the rest of the online services.

Of the 90 online users, 86 could be viewed on Instagram. Four had not posted anything during the period under consideration. With 2,678 posts, there was a 90% usage rate, meaning 9 out of 10 posts were on Instagram, and one was on another social network studied. The ratio of Stories to posts was 80%. Instagram Stories allows users to compile photos and videos for a slideshow that remains viewable for 24 hours. 1,027 locations were detected, which is a 38% detection rate. Most locations were detected with geotags or hashtags. The detection rate was highest for posts related to work at 70% and related to leisure at 63%. Additional clarification was achieved by manually researching terms and contexts. The overall recognition rate was 59%.

By comparison, the recognition rate for *flyinguwe* and *mamiseelen* was approx. 80%, and for Lena Gercke only approx. 40%. This circumstance could be related to the fact that the locations of *mamiseelen* and *flyinguwe* were known significantly more often. For *lenagercke*, few habits and hardly any background information could be derived from social media.

Examples of motion images:

Of *orchidee_2309*, 107 of 108 entries could be assigned geographically. Entries were made daily with the exception of one day, and daily routines were described. There was regular background information on family, occupation, and hobbies. The exact home address was on the letterhead.

fraui.jaci was very active with 187 entries. However, only 18% of the entries could be assigned geographically. Many entries were motivational sayings, selfies, showed family life or photos from different parties.

4.2   Facebook

42 onliners have a (known) Facebook user account. 33 of them did not publish an entry during the period under review. With 118 entries, however, the usage rate was only a very low 3.96%. 63 entries could be assigned to a geographical location. This corresponds to a recognition rate of 53.85%, which is even higher than that of Instagram. Geotags (56) and hashtags (55) led most often to the recognition of the location.

Content related to food or leisure had the highest recognition rate. Only 9 digital identities published at least one post on Facebook during the period under review. With 70 of 118 posts, *orchidee_2309* published just under 60% of all posts in the entire group.

4.3   Snapchat

For the onliners, 10 user accounts could be found on Snapchat. 6 digital identities did not publish any entries during the observation period. For 175 entries, the usage rate was 5.84%. 124 entries could not be assigned to a geographical location. This corresponds to a recognition rate of 29.14%. With the help of the snap map, significantly more entries could have been assigned to a location. However, this can only be used if the user's own location is revealed. For this reason, this function was not used. Most frequently, entries were assigned by the acquisition type "known" (34). This was followed by "visual" (20), "research" (12), "hashtag" (2) and "geotag" (1). Entries broke down into six general activities. The highest recognition rate is 87% by "work" with 23 entries and 20 recognised locations. "Leisure" follows far behind with 73 entries and 18 recognised locations (25% detection rate), other with 64 entries and 11 recognised locations (17%), and "eating" with 9 entries and one recognised location (11%).

"Shopping" (4 entries) and "spreading opinion" (2 entries) each could not be geographically assigned. Only the entries of the general activity "leisure" (17) and "eating" (1) were published in company. However, one location could not be identified. Of the total 124 locations that remained unknown, 5 could be identified based on the chronology of their publications (4.03% potential recognition rate) and 21 could be identified as residential addresses (16.94%) via the police information systems. This results in a potential recognition rate of 20.97%. This could increase the detection rate from 29.14% to 44%.

5   Discussion

In this research, 90 digital identities were to be observed over a period of 20 days in five online services with different objectives and user activities. After 20 days, it can be seen that the data from the 90 observed digital identities is suitable for drawing a movement profile for each digital identity.

However, it became clear that it is sometimes not always easy to track the digital identities even throughout the social media, as deletion or renaming of the respective user account is possible at any time and the user's subscribers are not informed about this either.

The snap map seems to be a tried and tested means of being able to locate users around the clock. There does not seem to be too strong an encroachment on fundamental rights despite the requirement of connecting two user accounts as "friends", as no verification of personal data is carried out and impersonating an acquaintance or friend of the suspect is usually not necessary. Snapchat user accounts are regularly publicly displayed on profiles of other online services. However, the Snap Map was not used for research because one's location must be disclosed in order for another user's location to be viewed. For the same reason, the use of the Snap Map could also be waived in criminal proceedings.

For some posts, it was questionable whether the publishing time also corresponded to the time the photo or video was taken. When publishing a post or story, either the camera of the end device or the album of the end device can be used. Consequently, there is always the possibility of using old files. Clues for matching the time of recording and the time of publication could often be provided here by lighting conditions and clothing. In addition, it was also questionable how meaningful a geotag or hashtag can be if the online services allow a location-independent selection of locations when selecting the geotag and any combination of characters and digits can be placed after a hashtag (#).

6   Conclusion

The data collection focused on the date and time of publication and the location. Other factors of great relevance were the activity and the accompanying persons. These data can be used to determine when a person was with whom at which location and what they were doing there. If these data are supplemented by several similar data, a movement profile emerges. If there are enough companions, a social profile can also be created. The work also showed that geotags generally provide a good indication of the location of the digital identity. However, other capture types such as information about the environment can also provide a good hit rate.

Instagram turned out to be the leading medium and is excellent for motion profiles. With Snapchat, a lot of potential could be seen especially in the Snap Map. However, some social online services such as YouTube and Twitter also turned out to be unsuitable. With the increasing digitisation of everyday life, more and more offenders are also moving in the digital world. Law enforcement agencies must not lose touch here.

7   References

Kroll, O.; Schwarz, U. (2001): Kriminalistik-SKRIPT. Die kriminalistische Fallanalyse. Studienbriefe für die Fachhochschulausbildung der Polizei. Kriminalistik 55 (3), pp. 215–216.

Labudde, D. (2018): Gruppen in Sozialen Netzwerken - Organisation und Dynamik. In: Honekamp, W.; Bug, S. (Hg.), Polizei-Informatik 2018. Remscheid: Rediroma Verlag, pp. 8–21.

Morrow, J. E. (2006): Open Source Intelligence. U.S. Army Field Manual Interim FMI 2-22.9. Hg. v. Headquarters, Department of the Army. US Army. Washington, DC. Retrieved on 31.03.2022 from https://fas.org/irp/doddir/army/fmi2-22-9.pdf.

Scheffler, H. (2014): Soziale Medien. Einführung in das Thema aus Sicht der Marktforschung. In: König, C.; Stahl, M.; Wiegand, E. (Hg.), Soziale Medien. Gegenstand und Instrument der Forschung. Wiesbaden: Springer VS, pp. 13–27.

Tesche, Y. (2018): Digitale Nutzung in Deutschland 2018. Abbildung der aktuellen digitalen Mediennutzung in Deutschland und Darstellung möglicher Trends, sowie Analyse des grundsätzlichen Verständnisses von Digitalisierung. Hg. v. Bundesverband Digitale Wirtschaft (BVDW) e.V. Düsseldorf. Retrieved on 31.03.2022 from https://www.bvdw.org/fileadmin/user_upload/BVDW_Marktforschung_Digitale_Nutzung_in_Deutschland_2018.pdf.

Wehlte, T. S. (2016): Open Source Intelligence. In: Honekamp, W.; Mielke, J. (Hg.), Polizei-Informatik 2016. Remscheid: Rediroma Verlag, pp. 115–122.

Welker, M.; Kloß, A. (2014): Soziale Medien als Gegenstand und Instrument sozial-
wissenschaftlicher Forschung. In: König, C.; Stahl, M.; Wiegand, E. (Hg.), So-
ziale Medien. Gegenstand und Instrument der Forschung. Wiesbaden: Springer
VS, pp. 29–51.