Instagram Activity Analysis: GDPR, JSON & Me: Research Proposal

Ryan Traviss

May 2020

1 Motivation

I want to discover what Instagram could learn about me if they analysed my personal data in depth. I am specifically interested in what my usage trends reveal from the data I know they hold. For example, could they work out when I catch the train to and from school purely from the timestamps on when I "like" photos? Or can they determine when I have just got back from a Scout camp based on a spike in the number of "followers" I have? Or are my usage habits just too random?

I am quite comfortable with my usage and feel like Instagram adds value to my life by allowing me to keep in touch with friends both new and old. Society often has privacy concerns about the data social media sites such as Instagram collect and a by-product of carrying out the above is seeing all the data they hold on me. Is this what I expected to find?

I propose to focus this project on Instagram as I know I use it regularly and have done since mid-2017 when I created my account. I considered Facebook but I mostly only browse less frequently and don't interact with posts much. I use Snapchat more to communicate with individuals directly similar to WhatsApp or text messages instead of sharing posts and seeing what others share like I do on Instagram.

2 Background

The General Data Protection Regulation (GDPR) [1] was written to give clear legal rights to people whose data is being used by other people or organisations. It is a European Union(EU) Regulation implemented in the UK by the Data Protection Act 2018[2] and therefore stayed in force after the UK left the EU. Before GDPR, across the EU, data protection laws varied and the penalty here in the UK was capped at £500,000[3] which was quite low. This research is only possible through the use of Subject Access Requests (SAR), commonly referred to as the "Right of Access". Basically, anyone can request an organization that holds data on them to give them a copy of it all which they must usually do within a month. There is no specific format the request must be made under [4]. I made a request each to Facebook, Instagram and Google on the 27th April 2020 which they complied with the same day. I further explain why I am only using my data due to GDPR in section 6.

Instagram is a social media platform, owned by Facebook, where you follow other people and share pictures with your followers to keep in touch. Social media sites often use lots of terminology but sometimes in conflicting ways and therefore follows a list of how I would use these key words to describe Instagram in particular.

- A "user" is simply a person who has an Instagram account. Eg. Ryan Traviss
- A "username" identifies an account to other users and always begins with an @ symbol. Eg. @exetermathssch
- A user can create a "post" which must include between 1 & 10 pictures or videos. I have made 51 posts at the time of writing.
- A post has a "caption" which is added by the user making the post, usually a poor attempt at a pun, and can be edited after posting. Eg. "We're all one big IKEA family!"
- Based on privacy settings, other users can then add "comments" to a post which consist of text or "like" the post by clicking a heart which shows they enjoyed it. "comments" can also be "liked". I have liked 8242 posts and 223 comments up until I downloaded my data on 2020-04-27.

- A user can “follow” another user meaning they see their posts if their privacy settings allow or they accept a “follow request” and these are referred to as “connections” in the data. I have 361 followers and I follow 537 accounts at time of writing.
- A user can “message” another user which is similar to sending a text message but can include pictures and other media.

It is convention to like all posts you see from your friends regardless of if you actually enjoyed them[5] and thus likes are a good approximation of usage.

In response to my requests, I received a nested file structure consisting of JavaScript Object Notation (JSON) files [6]. A JSON file consists of objects which are name/value pairs inside curly brackets such as ”language_code”: ”en” where the name is on the left and the value is on the right. There are also arrays which are an ordered list of values such as [”2020-03-24T10:31:20+00:00”, ”exetermathssch”] which is an array with 2 values inside as it is enclosed by square brackets. I further explain how I intend to manipulate the data in this format in the methodology section.

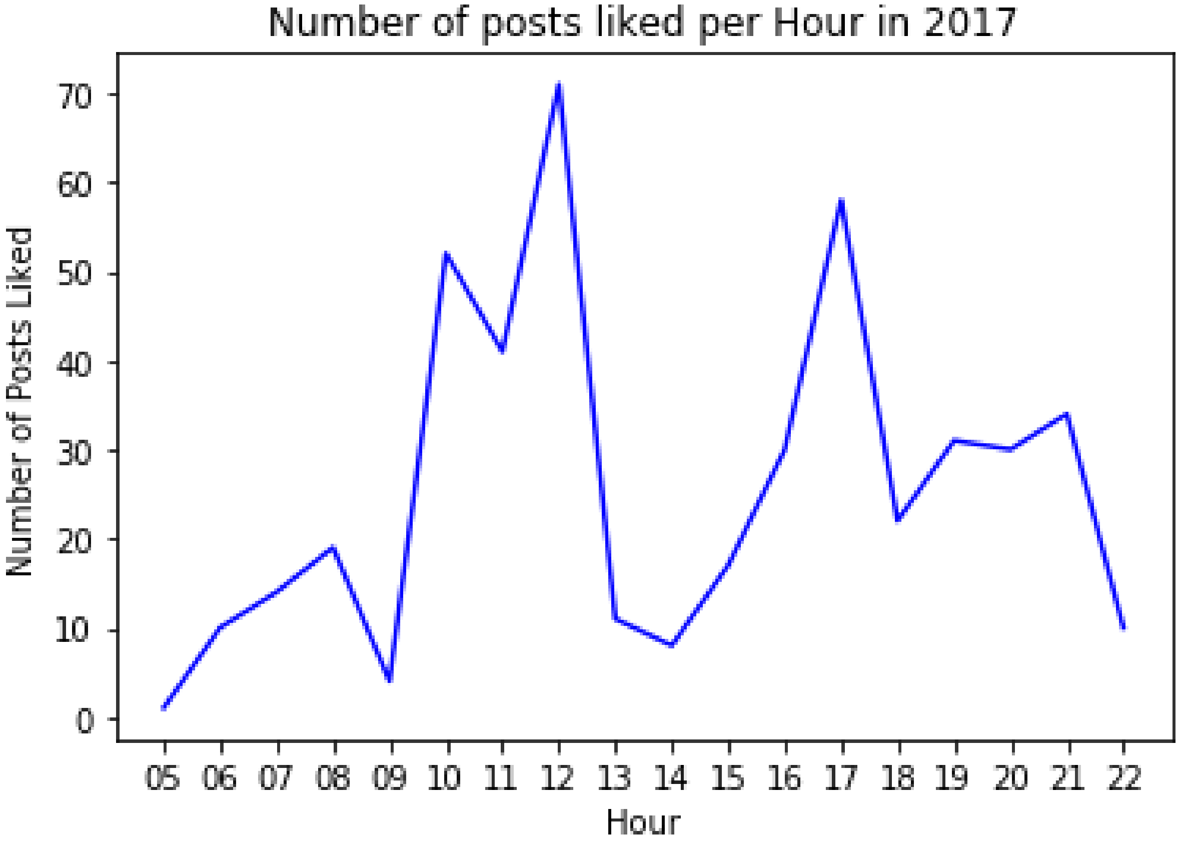
A key part of this project is manipulating dates and times of which there are many formats used to express them[7]. All dates and times I will be working with are in the format ISO 8601 [8] outlines. “2020-03-24T10:31:20+00:00” is an example which contains the date in the format year followed by month followed by day. This is followed by a T to represent the time portion and then hh:mm:ss for hours (in 24 hour format), minutes and seconds and then the time zone relative to Universal Coordinated Time (UCT) is the last part. This helpfully allows dates and times to be sorted alphabetically and be in order unlike the date format we normally use in the UK which is DD/MM/YYYY. Under this format, days of the week are numbered 1 through 7 where 1 is Monday and 7 is Sunday.

3 Tasks

1. To investigate my usage trends of Instagram using “likes”
   (a) By time
      i. On an hourly basis to identify peak times
      ii. On a daily basis to identify differences in weekend/weekday usage and on events such as Scout camps
      iii. On a monthly basis to identify wider trends such as between term-time and holiday usage
   (b) By user’s post I’ve liked: Is it possible to work out who my best friends are by this method? Or does this just reflect users who post more?
2. Aggregate Measure: To create an aggregate measure of usage based on “posts”, ”likes”, ”messages” and “comments”
   (a) Investigate trends over time and contrast to using only “likes” as above
3. Identify Best Friends: Investigate if you can work out who my best friends are through whose posts I’ve “commented” on
4. Compare Followers: To compare “followers” with events such as new schools and scout camps to determine the strength of this relation and thus if you could infer such events with only the “followers” data
5. Extension Graphs: Alternative methods of presenting the above information such as climate stripes for usage

4 Methodology

I intend to use Python 3 in order to build upon my existing knowledge with it and its use within data science[9]. I will use the built-in json library[10] which will convert the JSON file into a Python dictionary which is very similar but can be manipulated in Python. I will extract the quantities such as number of likes and times from this dictionary and then plot them using Matplotlib[11]. I intend to use the Object Orientated Programming paradigm[12] to reduce code duplication and ensure maintainability. I will be writing this in the Spyder IDE[13] due to it including many libraries I may use by default. Examples of these libraries include NumPy[14],which includes many powerful tools for efficiently manipulating large amounts of data, and datetime[15] which includes functions for manipulating dates & times such as converting a date into a day of the week.

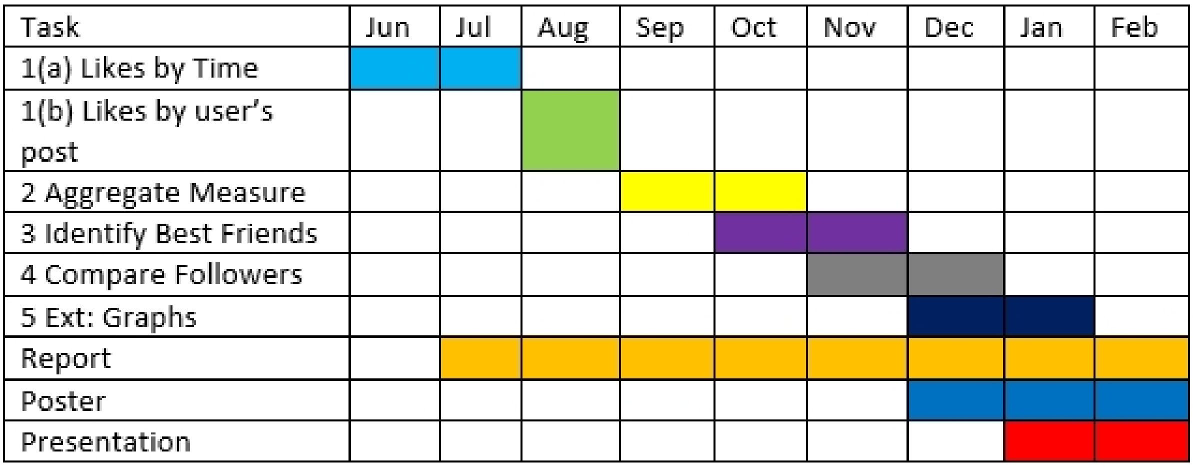


Above is an example graph that I have produced during the prototyping phase. The numbers along the x-axis represent hours on a 24 hour clock so 05 means 5AM and 22 means 10PM and the y-axis has the total number of posts liked within that hour during 2017. It is not within the scope of this proposal to conduct any analysis which I will do during the project. I intend to use train times, old school timetables and picture metadata to help explain possible reasons for trends I find and identify anomalies.

For working out if you can identify my close friends, I will write down, on paper, my close friends and then sort them into different ”levels” of friendship based on how much I trust them. This can be compared to the users with the most number of post & comments I've liked to see if this is an accurate way to identify my close friends. I will combine the likes count of people who have multiple accounts but make it clear when I have done this. This is a subjective and qualitative approach and I will evaluate this further in the report.

I am writing this document in LaTeX[16] and I will also use it to write my report & poster. This will allow me to develop my ability to produce professional scientific documents.

5 Timeline



Task	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
1(a) Likes by Time	■	■							
1(b) Likes by user's post			■						
2 Aggregate Measure				■	■				
3 Identify Best Friends					■	■			
4 Compare Followers						■	■		
5 Ext: Graphs							■	■	
Report		■	■	■	■	■	■	■	■
Poster							■	■	■
Presentation								■	■

6 Ethics & Bias

I plan on only using my own data in part to avoid ethical issues. I will not publish any messages I have sent/received and will be mindful of other people's reasonable expectation of privacy throughout. My Instagram username will not be published to preserve my privacy.

To prevent me altering my behaviour due to carrying out this research, I will work only using the data I downloaded on the 2020/04/27 prior to commencement of this project. At the moment, due to current restrictions in place to prevent the spread of COVID-19, I am also not meeting new people or going on Scout Camps so my behaviour is clearly going to be different than it would usually be.

I considered using other people's data to allow comparisons to be made and to see if correlations that are true for me are also true for them. However under GDPR, my lawful basis for processing[17] would be consent meaning they could, at any point, ask me to delete their data under the "Right of Erasure"[18] which I would then be obligated to do or potentially face legal penalties which are quite high[19]. Furthermore if I used data from my friends at school, they would be children whose rights are specifically highlighted in GDPR[20].

References

[1] *General Data Protection Regulation - Recital 63 - Right of Access.* Apr. 27, 2020. URL: `https://gdpr-info.eu/recitals/no-63/`.

[2] *Data Protection Act 2018 - PART 4 - CHAPTER 3 - Rights - Section 94.* Apr. 27, 2020. URL: `http://www.legislation.gov.uk/ukpga/2018/12/section/94`.

[3] *Information Commissioner's guidance about the issue of monetary penalties prepared and issued under section 55C (1) of the Data Protection Act 1998.* May 26, 2020. URL: `https://ico.org.uk/media/for-organisations/documents/1043720/ico-guidance-on-monetary-penalties.pdf`.

[4] *(Guide to the General Data Protection Regulation).* Apr. 27, 2020. URL: `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/`.

[5] Jin Yea Jang, Kyungsik Han, and Dongwon Lee. "No Reciprocity in "Liking" Photos: Analyzing Like Activities in Instagram". In: *Proceedings of the 26th ACM Conference on Hypertext Social Media.* HT '15. Guzelyurt, Northern Cyprus: Association for Computing Machinery, May 5, 2020, pp. 273–282. ISBN: 9781450333955. DOI: `10.1145/2700171.2791043`. URL: `https://doi.org/10.1145/2700171.2791043`.

[6] *Introducing JSON.* May 2, 2020. URL: `https://www.json.org/json-en.html`.

[7] *xkcd: ISO 8601.* May 2, 2020. URL: `https://xkcd.com/1179/`.

[8] *ISO 8601-1:2019 Date and time — Representations for information interchange — Part 1: Basic rules.* May 2, 2020. URL: `https://www.iso.org/standard/70907.html`.

[9] K. J. Millman and M. Aivazis. “Python for Scientists and Engineers”. In: *Computing in Science Engineering* 13.2 (May 5, 2020), pp. 9–12.

[10] *json — JSON encoder and decoder.* May 2, 2020. URL: https://docs.python.org/3/library/json.html.

[11] *Matplotlib: Visualization with Python.* May 5, 2020. URL: https://matplotlib.org/index.html.

[12] Peter Wegner. “Concepts and Paradigms of Object-Oriented Programming”. In: *SIGPLAN OOPS Mess.* 1.1 (May 26, 2020), pp. 7–87. ISSN: 1055-6400. DOI: 10.1145/382192.383004. URL: https://doi.org/10.1145/382192.383004.

[13] *Spyder: The Scientific Python Development Environment — Documentation.* May 5, 2020. URL: https://docs.spyder-ide.org/.

[14] *NumPy.* May 26, 2020. URL: https://numpy.org/.

[15] *datetime — Basic date and time types.* May 26, 2020. URL: https://docs.python.org/3/library/datetime.html.

[16] *LaTeX – A document preparation system.* May 26, 2020. URL: https://www.latex-project.org/.

[17] *Lawful basis for processing.* May 17, 2020. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/.

[18] *Right to erasure.* May 17, 2020. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-erasure/.

[19] *Penalties.* May 17, 2020. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-law-enforcement-processing/penalties/.

[20] *Children and the GDPR.* May 6, 2020. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/children-and-the-gdpr/.