Analyzing Instagram Post Frequency in Major US Cities

Neil Joshi, Duke Nguyen, Brandon Walsh
njoshi2@umbc.edu, du2@umbc.edu, brando12@umbc.edu

December 1st, 2015

Department of Computer Science and Electrical Engineering,
University of Maryland Baltimore County

Abstract

Social media offers a vast amount of data that can be analyzed in order to predict trends. Marketing agencies and businesses can use theses predictions in order to better engage their social media follower base. Post data gathered over a 1 week for 4 popular cities in the US and was used in order to determine the most frequent posting times. Regression with high polynomial degrees helped to determine popular posting times. The resulting statistics show weekly social media trends and popular posting times throughout the week.

Key Words - Social media, Instagram, Frequency Distribution, Polynomial Regression

1.0 Introduction

1.1 Mission Statement

Instagram is one of the most widely known social media platforms. It also provides a free API service with a fairly large limit. Members on our team had significant experience with marketing and social media. Analyzing social media trends is an effective way to learn how people think and function. Social media in conjunction with numerical computation can help us analyze trends and make informed predictions.We want to be able to determine when people are most active with social media posts. Using the data we can make many inferences that relate to marketing, business, society, and psychology.

1.2 Popular Works

There are many software and marketing companies working towards gaining insight from social media. A startup company called minter.io uses instagram analytics in order to help their users make more informed decisions when engaging with their Instagram followers.

1.3 Related Work

A study by coschedule.com used social media engagment on multiple platforms in order to find the best times to post. They found that Instagram audiences are more engaged throughout the week. The article by Julie Neidlinger also provides statistics for other important studies. According to TrackMaven, a digital marketing software, Instagram engagement stays consistent throughout the week, with slight spikes on Mondays and a slight dip on Sundays. The best time to post was during work hours, an average of 22.5 per 100 Instagram followers interacted with videos posted by popular companies. During off work hours these numbers increased to 33.4 hours per 1000 followers interacting with the content. Our methodologies differentiate because we are studying user post frequency rather than user engagement. This will lead to different results that must be interpreted differently.

1.4 Advantages of Our Work

Most research projects with similar target goals provide very high level data and do not educate the reader on how the results were achieved. Many neglect to see the power that lies in social media data. Being students with numerical computation experience, we are able to find results based off of large data sets we gather.

1.5 Contributions to Society

This work can help businesses know when and where to market their product. It could also be used for the average consumer, to find crowded city times. Specifically for Instagram users, people can discover popular times to post in order to receive the most attention. Overall, it's an excellent marketing tool and future tool.

2.0 Motivation

2.1 Problems Solved

Major cities can be overcrowded on certain days and times throughout the week. We want to discover the times an individual can avoid the crowd. We also want to use the data to find popular times marketing and social media agencies should post and engage with their follower base.

2.2 Related Problems

One major inspiration was the new Google feature which shows how crowded a business location will be at a certain time on a certain day. This information is used to help people avoid crowded restaurants, etc. This paper takes a spin on this concept to benefit marketers instead of actual consumers. Although only Instagram data doesn't contain nearly as much information as other sources, it is easy to access and has an API that is easy to use.

2.3 Problem Importance

Scraping large amounts of instagram posts take a long time. Instagram also has request limits that total 5000 posts per hour. We must also account for the fact that Instagram doesn't let you scrape further than one month in the past. Once we scrape the data, it can be used to make predictions about our society. Since a significant portion of our society uses Instagram the data also provides very much significance.

3.0 Methods

3.1 Novelty of Approach

Approaching this problem involved heavy preparation, as fetching current data from Instagram comes with limitations.It also comes with many benefits, as the data is expandable and open to further areas of research. Instagram is a fairly new platform and the Instagram API is even newer. With the constant increase in Instagram users it is the perfect opportunity to use Big Data in order to make predictions. We chose to use Instagram versus other social media platforms because it has a location based API and is more personal in terms of user interaction. Instagram posts focus on the actual user rather than news such as platforms like Twitter.

3.2 Specific Explanation of Code

Instagram's API allows users to make batch requests for posts. Each batch requests provided returns approximately 80 to 100 posts. In order to scrape thousands of posts we took the time stamp from the oldest post in each batch and made another API request at that particular starting time. The Pseudocode is shown below as a while loop:

```
while endTime =! currentTime:
        currentTime = lastpost.time
```

Once the raw unix timestamps are scraped from Instagram, they can be converted to a frequency distribution table by hour for each day the week. This is done by first converting the Unix time from the epoch, which is January 1st, 1970, into standard UTC, and then adjust it based on timezone.The list of times are then grouped by the hour of each day, and the frequencies are counted. The frequency data is stored in a text file, as it is important and concise information extracted from the large raw data file.

The next step is to take the frequency data and analyze it. The Python library, Numpy, is used to take the frequency data per hour and perform an nth degree polynomial regression to generate n+1 coefficients that make up a curve that fits the data.

Example: Baltimore on Thursday
Hour: x = [0, 1, 2, 3, ..., 23]
 # Posts: y = [101, 52, 19, ...]
5th Order Polynomial: y = -0.000542x^5 + 0.044805x^4 - 1.377944x^3 + 18.117957x^2 - 80.258757x + 110.910522

These curves are drawn along with plotted data points using the Python library, matplotlib, which generates the graphs described in the results.

3.3 Mathematical or Formal Issues of Techniques

Several tests were needed to determine the degree of polynomial to use for the curve fitting of the frequency data. A small order polynomial would under-fit the data, missing notable patterns in the data, but a large order polynomial would over-fit the data, creating jaring curves with sharp edges and several anomalies.

3.4 Parameters and Important Decisions

The most important parameters are a lat/lng and a unix timestamp. Data was scraped into large text files and analyzed using our algorithms. By default, a 10th order polynomial is used for per-day data, and a 50th order polynomial is used for data consisting of the whole week, which has several more curves to account for.

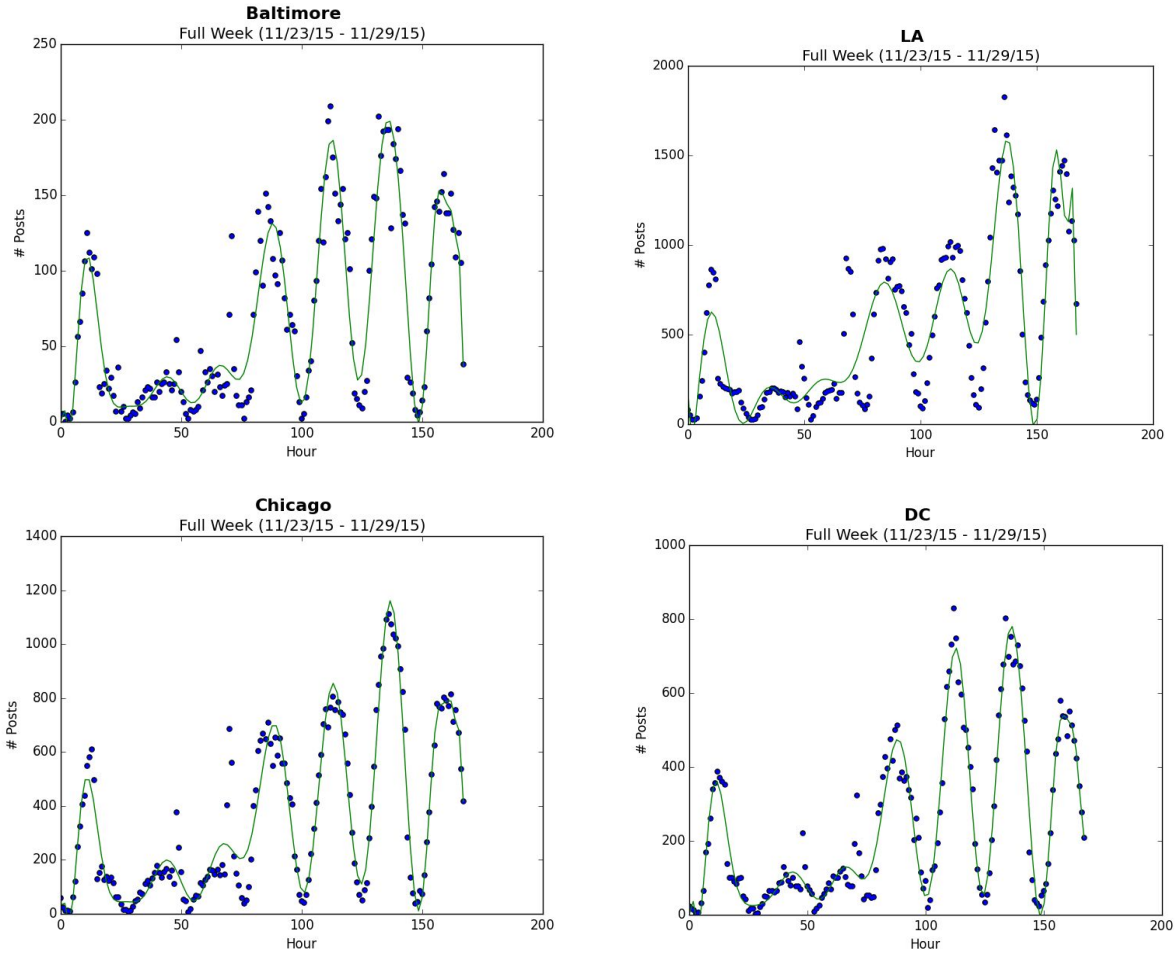
3.5 Plan of Action

Our custom instagram post prediction system allows us to make very concise, accurate predictions. We will scrape every instagram post for a 7 day period in 3 popular cities. We will then use the data in order to form lines of regression with high degree polynomials. The large amounts of data, 100k+ posts will allows us to get very accurate predictions.

4.0 Results

Data was gathered from Baltimore, Chicago, DC, and LA

4.1 Weekly Graphs Comparison



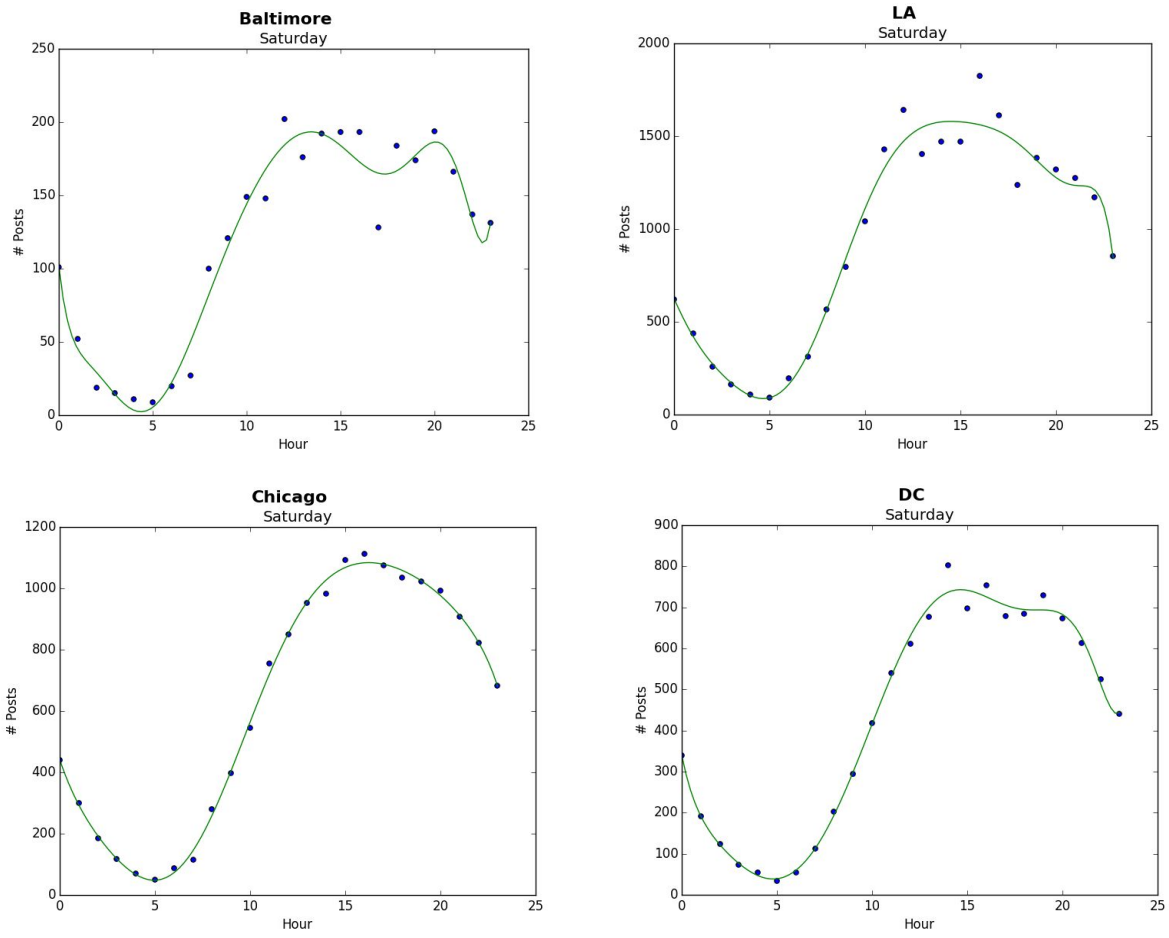
The 4 graphs above show regression data for Chicago, Baltimore, DC, and LA. The x-axis represents 168 total hours (1 week) and the y-axis represents the total number of posts. Visual analysis of the data shows that each city has a very similar posting frequency with the exception of Chicago and LA posts being much more popular on Saturday. The total posts for each Chicago, Baltimore, DC, and LA total approximately 20000, 100000, 65000, and 143000. The differentiation is due to the population density of the city as the same distance area was analyzed for each.

These weekly graphs lets us infer the popular posting times are the weekend and peak at around 5pm. The graphs above were scraped over the Thanksgiving Holiday, possibly making Thursday

skewed from the average non holiday weekday. Tuesday and Wednesday were significantly less active, with very minimal posts. Also important to note is the much higher frequency of posts on the weekend in LA, compared to the other cities.

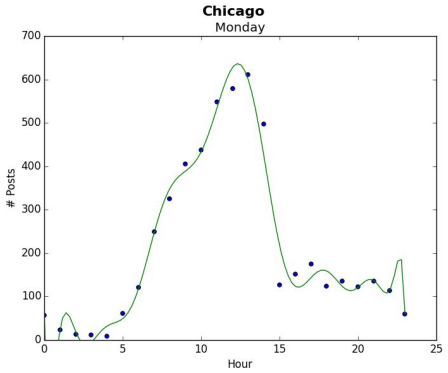
4.2 Daily Graphs City Comparison

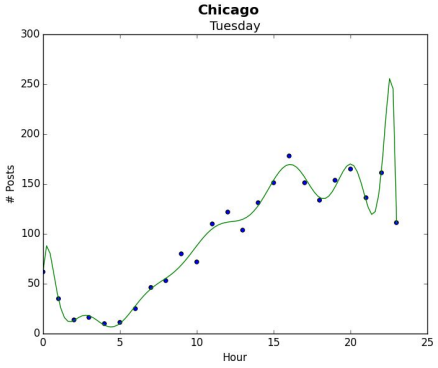
Saturday is used as an example

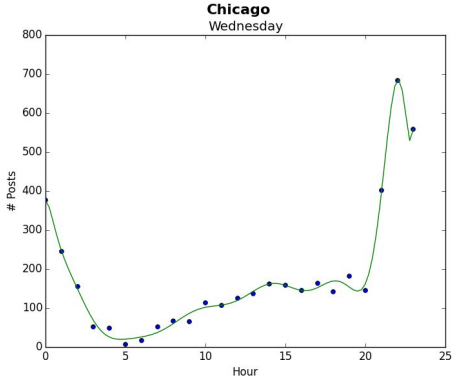


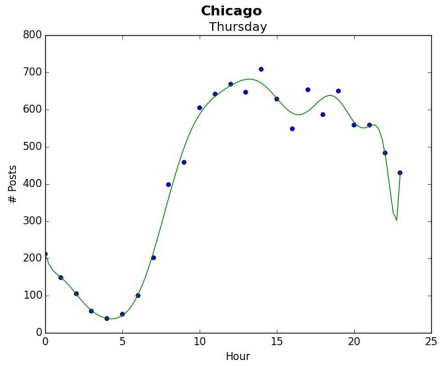
Each city shows a very similar trend on saturday, majority of the posts occur around 3-5pm. This shows that each city has similar Instagram users. All graphs shown use 24 data points and 10th degree polynomials.

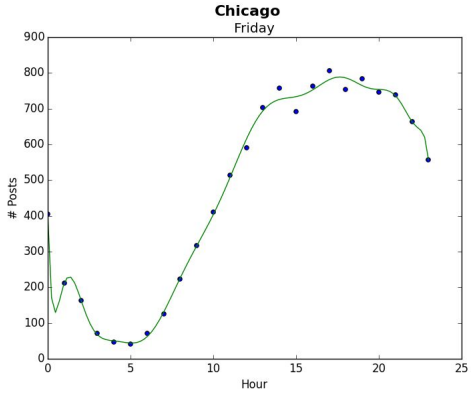
4.3 Daily Graphs for Chicago

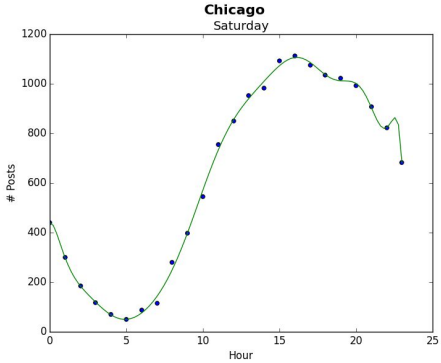


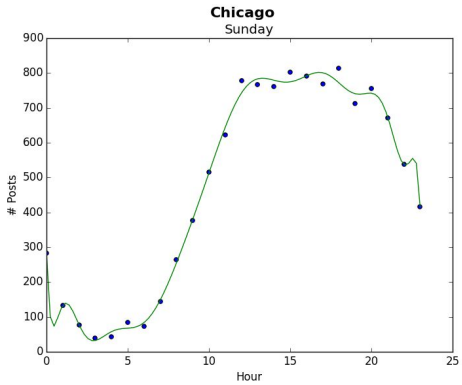








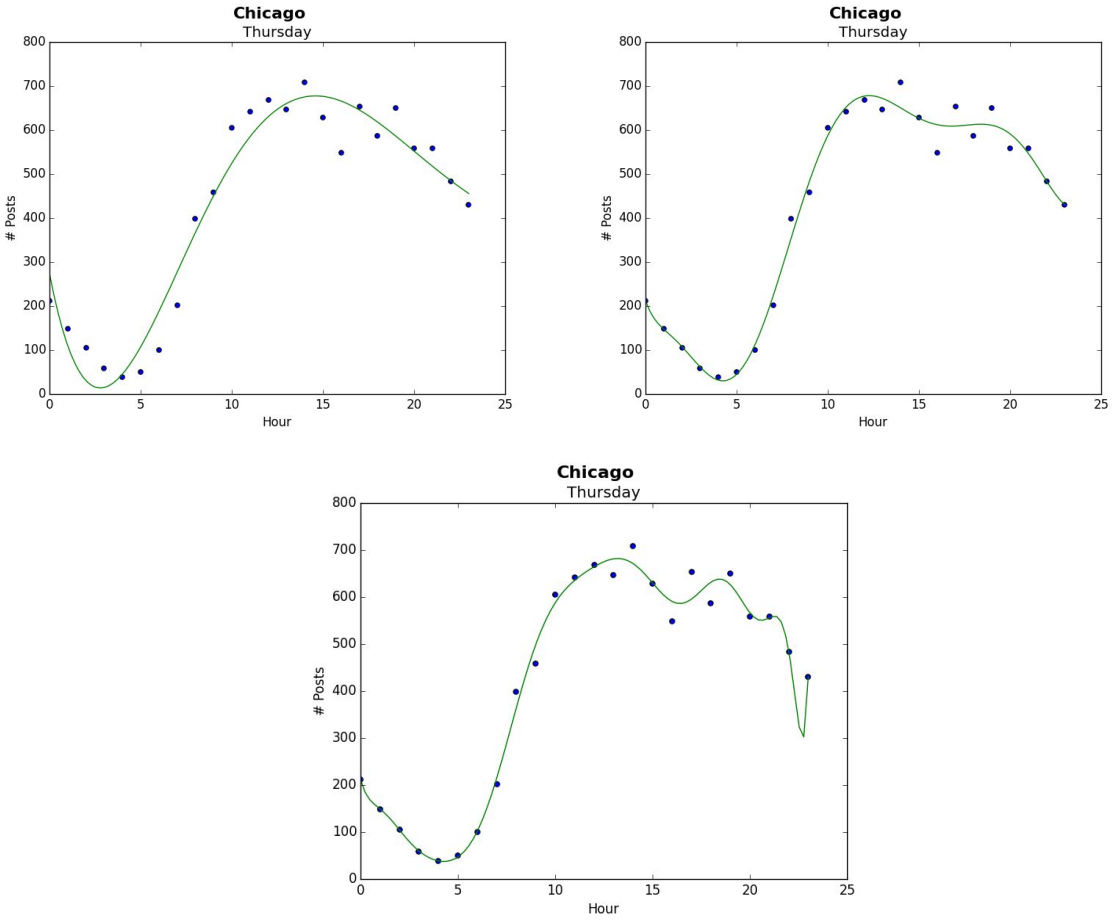




The graphs above show every day of the week for Chicago. Each graph uses regression with 15th degree polynomials. The weekend provides a similar trend curve but each day during the week offers useful insight. For example, it is likely that Monday has such a big jump due to people posting their images from the weekend.

4.4 Comparison of Different Degree Polynomial Curve Fits

Chicago on Thursday is used as an Example



The graphs above exemplify the difference between using a 5, 10, and 15 order polynomial curve to fit the data points. Using a 5th degree polynomial creates a very broad curve that misses important details during the day, like lunch time, etc. The 15th degree polynomial demonstrates the troubles with using very high order polynomials, in that the curve tries to fit the data so much that it deviates from the broad trend, as shown between hours 22 and 23. The 10th degree polynomials shows a healthy compromise, which is why it is the default degree for most of our daily graphs.

4.6 Summary of Results

The figure below shows the local maximums of each cities weekly graph.

	Monday	Tuesday	Wed	Thursday	Friday	Saturday	Sunday
Baltimore	12:30pm	8:00pm	12:00pm	2:00pm	3:00pm	2:00pm and 8:00pm	2:00pm
Chicago	12:00pm	5:00pm	11:00pm	12:30pm	6:00pm	4:00pm	2:00pm
DC	12:30pm	4:00pm	12:00pm	3:00pm	4:00pm	2:00pm	2:00pm
LA	10:00am	12:30pm	9:00pm	11:00am	8:00pm	2:00pm	5:30pm

Since the data was scraped over a holiday week there is proof that the data is accurate. For example, the Wednesday column has very late times for a week day. This is because the night before Thanksgiving in a popular time to go out.

It's also important to note that Wednesday is the most inactive day on Instagram. There also appears to be a large amount of Instagram posts on Monday during lunchtime, which could be explained by people posting things from the weekend.

5.0 Discussion

5.1 Summary

A social media prediction based system has been created using Instagram and numerical methods. The results found have been substantial and have conformed with our predictions. Social media posting has become such an integral part of people's lives that you are able to use it in order to predict trends in real time. The data used provides evidence to when people are most socially active on Instagram.

5.2 Achievements and Claims

Using 100k+ total posts we are able to make substantial predictions as to popular times people post to instagram. Saturday is the most popular day of the week for Instagram. 4pm was estimated to be the most popular time of the day to post, although this depended on the day heavily.

5.3 Areas of Improvement

The most glaring issue with this research is the timeframe used, which was Thanksgiving week, was not a very representative week in relation to the average week. This was a limitation of Instagram, as scraping data for more than a week would take much longer. Not only did Thanksgiving and Black Friday skew our results, it also made scraping much slower, because there was a larger volume of Instagram posts than normal. Large cities such as NY also caused the scraper to break due to the large amount of posts, 1 million+ for only a week.

5.4 Future Research and Opportunities

There are many possible directions to take with this project, as the scope is narrowed down to be appropriate as a course project. One option would be to create a software that offers marketing predictions to social media companies and small businesses. We can also expand our data set and compare data from more major cities. We can also decrease the time interval from 60 min to 10 min and increase the polynomial count, and do deeper analysis. Lastly, we could harness the remaining Instagram data and expand the project much further, possibly by performing social psychological studies.

6.0 Acknowledgements

Instagram API Endpoints. (2015). Retrieved October 9, 2015, from https://instagram.com/developer/endpoints/
This document contains relevant information about Instagram's developer APIs.

Press News. (2015). Retrieved October 9, 2015, from https://instagram.com/press/
This page shows some first-party statistics straight from Instagram. This is useful to show the state of the social media network.

Beese, J. (2015, January 22). 5 Insightful Instagram Statistics That You Should Know. Retrieved October 9, 2015, from http://sproutsocial.com/insights/5-instagram-stats/
This article shows some more statistics about Instagram. In particular, the page showcases statistics that a more useful for marketing agencies.

7.0 References

1. “Instagram API docs.” , https://instagram.com. Kevin Systrom, 2015, INSTAGRAM.Web.12/01/2015
2. “Python.” ,https://www.python.org/. Python Software Foundation. Web.12/01/2015
3. Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37
4. Hunter, J. D. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), IEEE COMPUTER SOC
5. “Facebook messenger.”,https://www.messenger.com. Facebook, 2015.Web.12/01/2015
6. “GitHub” ,https://github.com. 2015 Github, Inc.Web.12/01/2015
7. “Google docs”, https://www.google.com/docs/. Google, 2015.Web.12/01/2015