Success Dynamics in an Online Photo Sharing Platform

Which will be your most popular post?

Authored by

ZSOMBOR KOMAN

Advised by

ROBERTA SINATRA

and

ROSSANO SCHIFANELLA

Department of Mathematics and Its Applications
CENTRAL EUROPEAN UNIVERSITY

MAY 2018

Abstract

The thesis explores the relationship between the dynamics of performance and success in the case of an online photo sharing platform. The careers of photographers are studied in terms of peak achievements, as introduced in the field of science of success.[1] A fundamental and robust difference is observed as compared to previous results, multiple hypothesis are formulated towards explaining the phenomenon, initial tests are conducted and directions for further research are pointed out. Aesthetic scores generated through deep learning are used to assess the quality of the individual images.[2] Then matching methods[3] are applied to infer causal relationship between the aesthetic value of a photograph and the probability of positive social feedback received.

ACKNOWLEDGEMENTS

First of all I would like to thank my thesis advisor Roberta Sinatra for her giving me the opportunity to work on this project and for providing constant advice and feedback throughout the research process. I am also very grateful to the authors of the paper *Beautiful and damned. Combined effect of content quality and social ties on user engagement*, especially to Rossano Schifanella and Luca Aiello, who shared their insight in personal discussions as well. In addition to that, Rossano became my second advisor for the last steps of the process, helping me out whenever needed in the concretization of the results and the formulation of the thesis itself.

Besides these officially documented interactions I would like to express my thanks to Federico Musciotto for helping with the literature connected to matching methods, reading through the text, commenting on both the most abstract and the very down-to-earth issues.

Finally, I must express my gratitude for the background given by my family throughout the past decades, the importance of good soil cannot be overstated. Regarding the more recent years and the time spent on this thesis I must sincerely thank Zsuzsanna Lakatos, who managed to restore my mental stability in the hardest times of this process through methods ranging from reminding me that there is some ice cream in the fridge through learning to beat me in multiplayer strategy games or even forcing me to sleep when I was close to becoming a zombie[1].

[1] I know, it's pretty easy for me to become a Zombi, I only need to loose one letter from the short version of my first name to achieve that: many thanks for all the people who so often manage to make the dumb joke of pointing this out – the world would collapse without their support.

TABLE OF CONTENTS

Page

Abstract **i**

Acknowledgements **iii**

Introduction **1**

1 Theoretical framework **3**
- 1.1 Modeling success . . . 3
- 1.2 Modeling performance . . . 7
- 1.3 Causal inference based on observational data: matching methods . . . 8

2 Data and general statistics **13**
- 2.1 Success and performance on a photo sharing platform . . . 13
- 2.2 The sample and its structure . . . 15
- 2.3 Data field distributions . . . 17
- 2.4 Unique upload dates and defining albums . . . 18
- 2.5 Trends and normalization . . . 19
- 2.6 Correlations . . . 20

3 Results **23**
- 3.1 Random impact rule . . . 23
- 3.2 Possible causes, hypothesis and analysis . . . 25
  - 3.2.1 Album-based analysis . . . 26
  - 3.2.2 User profiling . . . 27
  - 3.2.3 Diversity . . . 28
- 3.3 Causality analysis . . . 28

Conclusion **33**

A Appendix **35**

Bibliography **39**

INTRODUCTION

Social media is becoming an organic part of our everyday lives. We spend a significant amount of time on these platforms. And often we ask ourselves: why does this or that post get so much attention, what do I have to do to get more likes, views or comment? Does performance underpin success? Or conversely, is success explained mostly by performance (quality of our posts, professionalism of photos taken)? In this thesis these and similar questions are in focus. The analyzed data is from an online social platform designed for photo sharing. The topic is approached from an empirical point of view using robust statistical methods and widely applied modeling schemes.

The main goal in this thesis is to analyze the careers of photographers who upload their work to Flickr. One possible approach is to try applying the methodologies used in the field of science of success,[1] observe the similarities and differences between the results and draw the needed conclusions as well as pointing out possible further directions. Progress in this direction is the first aim of this project.

The data used gives us an unusual advantage: we can make use of performance measures besides the ones describing success. In most of the targeted areas performance is very hard to measure or even define. Imagine, for example, the case of a scientific paper (which is by the way the most studied case in the field). We cannot really associate an objective performance measure with it. Who could decide how valuable a paper is? Of what quality does is it? How much innovation does it bring along and how much will it change the state of science and the way our world will look like in 10 years or a century?

But for photos we can use the results of computational aesthetics to define a beauty or quality score of each uploaded photo based on a machine learning algorithm (which was already developed and used in previous research being an attempt to capture the consensus or average of human votes about the aesthetic value of photographs).[2]

Although such an estimator might be criticized for lacking an intrinsic definition of beauty, it is hard to argue with the fact that the score obtained this way is objective, i.e. it is independent from the person of the author, its social position, etc. So in addition to analyzing the evolution of a career from the perspective of performance measures, we have the possibility to find out more about the possible causal ties between performance and success.

For the assessment of causal interactions findings from the field of matching methods will be used. Testing the existence of such ties using mathematically well-backed methods is the second

central theme of the research project.

To sum up, the two main aims of the thesis are the following: to better understand the success and performance dynamics of the careers of photographers on the studied online platform and to conduct randomized experiments on observational data for inferring causal connections between quality and popularity.

The first chapter will give an overview of the three fields on which this project is built: science of success, computational aesthetics and matching methods. The second chapter will present some of the basic preprocessing steps and statistical features of the data. The third part will conclude by addressing the research questions: applying the presented methodologies to the data, comparing the found results with our expectations, stating then assessing some hypothesis and pointing out possibilities for further research.

CHAPTER 1

THEORETICAL FRAMEWORK

This chapter is meant to present the most relevant parts of the academic literature providing the theoretical background for the thesis. As presented in the introduction, the research is based on two separate streams: science of success and computational aesthetics. Besides these two areas a third section of this chapter will deal with matching methods, a field concerned with the statistical background of performing experiments on historical data. This will be used to create control and treatment groups rigorously and achieve unbiased results (similarly to randomized experiments, where selection is much better separated from treatment and its effects).

1.1 Modeling success

To talk about the science of success we must define what we mean by success in the first place. Most often it is considered a benefit or reward gained by outstanding deeds. Some example realizations of success can be popularity, revenue, impact or power. These can be measured for example in number of followers, views, likes, amount of dollars earned, assets under management, votes, etc.

These examples show two basic common characteristics of success. First, it is a collective phenomenon representing the judgment of the community regarding someone's achievement or ability. Secondly, although success is the result of such a complex social behavior, it can be proxied and quantified reasonably well through any of the enumerated example measures.

But what about the performance behind the success? It is the output of an individual (or potentially a team) and usually it is much harder to measure. In some cases it is easy to define performance metrics: the time it takes for someone to perform a task (example athletics) or the difficulty of the task formulated in terms of numbers (weight one can lift or push-ups one

can do). But in many interesting real life situations, although success can be easily measured, performance is very hard to express in terms of numbers. Just think of science: we have many metrics of success like citations, prizes, grants, but it is really hard to assess the quality of research without taking into account the response of the surrounding community in one way or another.

When considering things from this perspective and after understanding that performance is the output of the individual while success is based on the perception of the society, the question if the two correlate with each other becomes a relevant and non-trivial. One could say that in the case of sports, where performance is well-quantified, the highest performance will get the first prize for sure and thus success is well-aligned with the results presented. But if we consider the income as the measure of success, this relationship is not so trivial anymore: in most of the cases sports-persons with highest earnings receive a big portion of their benefits from advertisement, which is very much influenced by the fan-base of the given person.

Another interesting success metric can be popularity defined as the number of Google searches or Wikipedia page views (during a given time period). A recent study by Albert-László Barabási and Burcu Yucesoy shows that there is a strong relationship between the professional results of tennis players and their Wikipedia page views. They even came up with formula for predicting the future values of this metric during the player's active years and even after retirement:

$$W_M(t) = A\frac{Y(t)}{r(t)}V(t)n(t)e^{\Delta r(t)H(\Delta r)/r(t)} + C\frac{Y(t)}{r(t)}, \tag{1.1}$$

where $W_M(t)$ represents the momentary Wikipedia visits of the player, $r(t)$ symbolizes the rank of the player, $\Delta r(t)$ the difference between the ranks of the considered player and it's highest ranked opponent in the championship, $V(t)$ the value of the tournament considered, $n(t)$ the number of matches played by the given player on this event and $Y(t)$ denotes the length of the career.[4]

Although it seems to be true that performance and success are coupled, the connection between them is not as trivial as we might think and additionally the transformation which connects the two is far from being a simple linear one. As Sherwin Rosen puts it when analyzing the superstar phenomenon, i.e. concentration of rewards among a few individuals, there are two important factors leading to it: "first, a close connection between personal reward and the size of one's own market; and second a strong tendency for both market size and reward to be skewed toward the most talented people in the activity".[5] The local dynamics of preference decisions, the hard to perceive small differences in performance and the belief in the wisdom of the crowd all underpin and explain this global idea of skewness. These observations are expressed mathematically through the highly convex nature of the transformation mapping from the talent or quality domain into that of the associated rewards. In the end remarks of his work, Rosen points out that one of the process' catalyzers are the new facilities for communication, which can take this effect to the extremes through fast and closed feedback-loops (better performance leads to higher success which leads to higher market share thus even higher success, etc.). Expressing

this forecast in the 80's makes him one of the prophets of the pop culture defined by the internet and social media in our days.

Given these observations, our everyday experiences are not so surprising anymore: human performance is typically bounded by our physical capabilities but the success we can achieve is only bounded by the sizes of the potential markets. The processes generating success from performance are multiplicative and iterative in nature thus resulting in exponential mappings. This is how the popularity of the fastest runner can be orders of magnitudes higher than that of its competitors although their results only differ by a couple of percents.

We are maybe even less surprised by these dynamics if we consider the structure of the world wide web[6] which is well-explained by the Barabási-Albert model.[7] The generative dynamics is described by a multiplicative stochastic process which results in the Pareto-distribution of the node degrees:

$$p_i = c \cdot k_i, \tag{1.2}$$

where p_i is the probability of node i to receive a new connection and it is proportional to its degree k_i, i.e. the number of connections it already has (c is a normalization factor). This effect, known as preferential attachment, is in fact the network science equivalent of what we could express as "success breads success" in the popularity domain. Similar multiplicative processes resulting in Pareto-distributions in the most varied areas and measures are well-understood and documented in the literature.[8]

Besides the mentioned effects an interesting and relatively new insight comes from the study of scientific careers lead by Roberta Sinatra,[1] the law called the random impact rule being in the spotlight. In the case of scientific research it is extremely hard to define an objective and intrinsic performance metric, so this approach tries to understand some fundamental properties of the success dynamics without speaking about performance at all. Here the main quantity considered is c_{10}, the number of citations a paper receives in the first 10 years after publication (the fixed time interval is considered to make the measure independent of the publication date and the interval length is determined by the decay dynamics of citations). The first interesting finding illustrated in figure 1.1(a) is that no clustering of success is observable, meaning that the c_{10} of the articles just before and exactly after the top hit (c^*_{10}) are on average identical with the c_{10} averaged over the entire career. Formulating this in everyday terms we could say that scientists do not learn from their biggest success, they do not find the answer to the ultimate question and after that they become more cited.

The second finding only extends this result to the entire career: there is no statistically significant difference between the original careers of the scientists and the randomly reshuffled ones (shuffling means that the publication dates are reordered while keeping every other characteristic including the citations of the papers intact), when we consider the timing of their most successful paper as shown on figure 1.1(b)). Although there is a decay in the probability of success after 20 years, this is not due to less creativity or poorer performance, but simply the lack

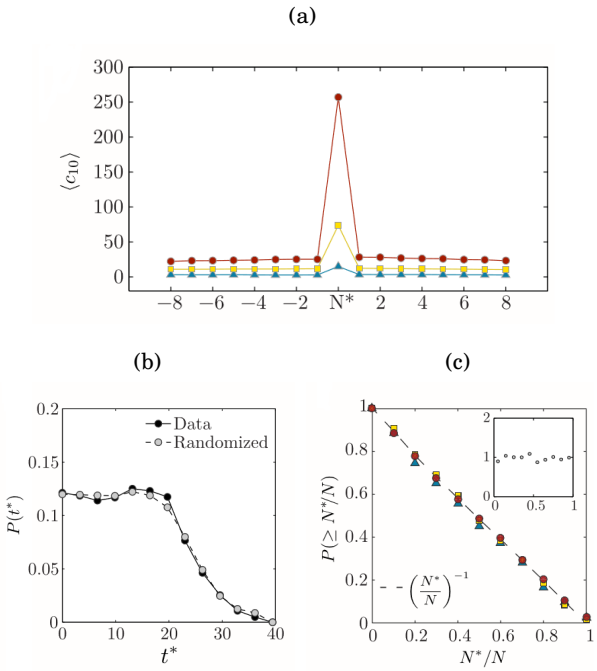

FIGURE 1.1. (a) No clustering of success. (b) Probability distribution of most successful paper as a function of time. (c) Inverse cumulative distribution of most successful paper as a function of rank. Figures taken from the original paper.[1]

of productivity (otherwise there would be significant differences when compared to the reshuffled version). Equivalently, the rank of the most successful publication (N^*) is uniformly distributed throughout the career of a scientist, which means that the cumulative distribution function of this indicator is linear (figure 1.1(c)). In plain English we could say: nobody knows when your best hit comes.

The random impact rule suggests a trivial null model: the success of each and every paper is drawn from the same distribution, thus the c_{10}^* of each researcher only depends on their productivity. But this would imply a success distribution which leads to lower slope of the productivity - maximal impact curve and the scientists having consistently high average impact would have divergent maximal impact (or conversely: there shouldn't exist scientists with consistently high average impact at all). Thus the first model needs to be rejected based on the data, but extending the model by assigning a quality score to each researcher is able to explain all the measurements correctly. This extended model describes scientific impact of a paper as being governed by the following stochastic law:

$$S = Q_i \cdot p, \tag{1.3}$$

where S is the impact (success), Q_i is the quality score associated with each scientist i and p is a value drawn from the common distribution for all scientist which is independent of both N (the length of the career) and Q_i (the quality or skill of the given researcher).

Besides the coupling between the performance and success of an individual, the interactions between individuals are also important towards building an understanding in this field. On one hand individuals might be able to learn from each other and increase their performance,[9] on the other hand network position is an important asset for increasing the market share mentioned by Rosen as well, which results in increasing success.[5]

This section was focused on showing the fundamental differences between success and performance, making the point that these differences make the study of their correlation interesting and showing the possibility of describing their coupling as well as understanding the stochastic dynamics of success alone when no associated performance metric is given.

1.2 Modeling performance

The other important scientific endeavor which will help us get an objective quality measure for photos is that of computational aesthetics.

In the first phases of this research area, assessing the aesthetic value of a photo meant defining hand-crafted features like spatial distribution of edges, distributions of colors, hues, contrast, measures of blurriness and focus and identifying low-level features. Even more than that one had to classify photos based on styles to identify if a blurry spot signifies a bad quality photo or is rather an artistic trait.[10]

Nowadays deep learning is the tool used to assess this problem. The input of the neural network consists of the encoded pixels of a photo. After this a convolutional layer transforms the photo into locations and intensities of features, e.g. there is a sharp edge in the top left corner of the image. These are then passed to a regular deep learning network.[11] To capture both the local and global characteristics of an image, two such convolutional network columns are created, one concentrating on the low level, the other one on the high level properties (trained on corresponding crops of the images). The two layers are then combined through a final network layer taking as input the outcomes of both columns and outputting an aesthetic score for the image, as presented in figure 1.2.

There are also possibilities to incorporate the photographs' style in the calculations. In the above cited article this is done by training an additional network to classify style based on a smaller subset of images being tagged with this ground truth data. Afterwards the output of this network is used as an additional feature for refining the results of the previous approach as figure 1.3 shows. This additional complexity significantly increases the accuracy of the classification.[11]

In the case of our dataset a pretrained object identification network is modified and fine-tuned for the purposes of aesthetic value assessment.[2]

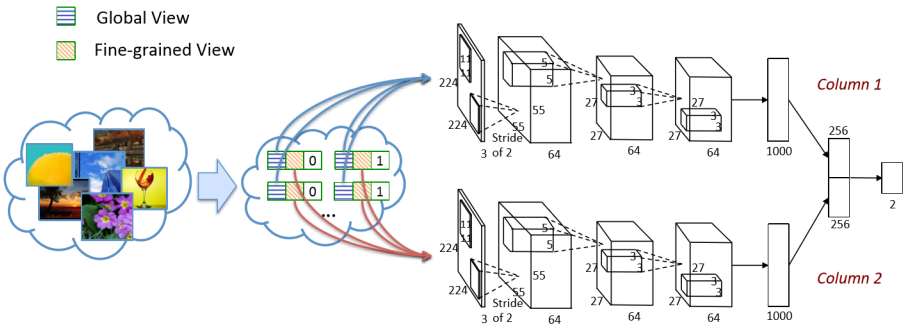

FIGURE 1.2. Two column convolutional neural network architecture learning both global and local features. Figure taken from.[11]

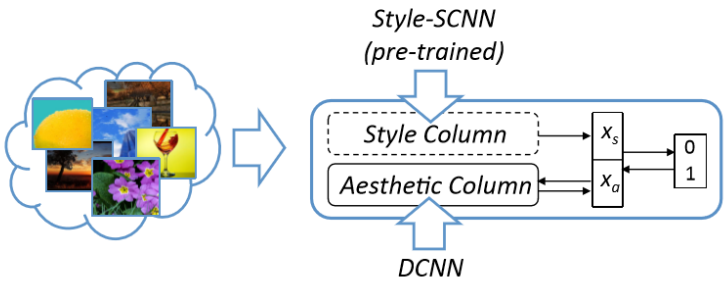

FIGURE 1.3. Learning style identification and incorporating this information in the process of aesthetic value assessment. Figure taken from.[11]

1.3 Causal inference based on observational data: matching methods

The third line of literature brought into our investigation is related to the toolkit of matching methods.[3] In the typical case of dealing with data and trying to prove causal relationships researchers face serious difficulties in avoiding bias. First of all, correlation of specific fields is not enough to deduce causal relationship between the two variables considered. The only valid way of testing causation is by doing random experiments, where the approach is to have treatment and control groups which have the same covariate structure, i.e. have the same joint distribution over all the variables which might have any effect on the outcome of the treatment.

Matching methods are designed to simulate or imitate such experiments based on data. The difficulty is caused by the selection of the treatment and control groups in an unbiased and balanced way while preserving enough data points to be able to evaluate the treatment effect with good statistical significance.

A matching method is a procedure which decreases the distance between the covariate

distributions of the treatment and control groups and thus reduces bias due to the covariates (selection bias). The distance which is to be minimized can be defined in several ways and it is often done on an individual-to-individual basis thus requiring the method to use this one-to-one distance to decrease the distances between the two distributions, e.g. trivially selecting pairs of individuals with minimal such distances (nearest neighbors).

The most widely used distance measure for these purposes is the Mahalanobis distance which works well for continuous variables and takes into account the correlations between the covariates as well.

Definition 1.1. Mahalanobis distance.

$$D_{ij} = \left(X_i - X_j\right)' \Sigma^{-1} \left(X_i - X_j\right),$$

where X_i is the covariate vector of the i-th individual and Σ is the covariance matrix of the considered variables.

In the case of high-dimensional data using such measures directly makes it very hard to do the selection without drastically decreasing the number of individuals taken into consideration. Reducing the number of participants in the experiment will reduce the statistical significance of our analysis, so we would like to avoid these approaches.

It turns out that we can find scores of reduced dimensionality on which we can apply similar minimization procedures to reduce bias while not loosing too much statistical significance.[12] Such scores are called balancing scores and defined as follows:

Definition 1.2. Balancing score.

$$b(X_i) \text{ balancing} \iff W_i \perp X_i \mid b(X_i),$$

where W_i is the treatment indicator and X_i the covariate vector.

Observe that the covariate vector itself is a balancing score by definition. But as described earlier we are interested in reducing dimensionality to realize matching without loosing too much statistical significance. We will show that there is a one-dimensional balancing score which moreover is the coarsest (can be expressed as a function of any balancing score).

Our candidate is the so called propensity score, which is the probability of getting treatment conditioned on the covariate values.

Definition 1.3. Propensity score:

$$p(X_i) = \mathbb{P}(W_i = 1 \mid X_i = x).$$

So first of all we need to show that this score is a balancing score as well.

Proof. Balancing property of the propensity score.

We want to show that

$$W_i \perp X_i \mid p(X_i) \iff \mathbb{P}(W_i = 1 \mid X_i, p(X_i)) = \mathbb{P}(W_i = 1 \mid p(X_i)).$$

For the proof we show that both sides of the equation are equal to the value of the propensity score in fact.

$$\mathbb{P}(W_i = 1 \mid X_i, p(X_i)) = \mathbb{P}(W_i = 1 \mid X_i) = p(X_i)$$

By observing that the information contained in the covariates and a function of them is the same as the information contained in the covariates themselves and using the definition of the propensity score.

$$\mathbb{P}(W_i = 1 \mid p(X_i)) = \mathbb{E}(W_i \mid p(X_i)) = \mathbb{E}(\mathbb{E}(W_i \mid X_i, p(X_i))|p(X_i)) =$$

$$= \mathbb{E}(p(X_i) \mid p(X_i)) = p(X_i),$$

where we used the law of total expectation and the definition of the propensity score again.

■

If the covariate distributions are balanced and they were defined such as to cover the range of variables which might be in any relationship with the possible outcome of the experiment, then the so called unconfoundedness assumption holds, i.e. the possible outcome of the treatment is independent from the treatment assignment given the value of the covariates:

$$W_i \perp Y_i(0), Y_i(1) \mid X_i,$$

where $Y_i(0)$ is the value of the outcome without treatment (in the control) while $Y_i(1)$ is the outcome with treatment. We want to be sure that the unconfoundness assumption is not being violated in the process of balancing based on any balancing score. Note that this assumption can never be guaranteed in real life situation, it usually holds only up to a certain degree, but we want to be sure not to make it worse through our balancing process.

Proof. Unconfoundedness given a balancing score.

We want to show that

$$W_i \perp Y_i(0), Y_i(1) \mid b(X_i) \iff \mathbb{P}(W_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = \mathbb{P}(W_i = 1 \mid b(X_i)).$$

For the proof we transform the left hand side to be equal to the right hand side.

$$\mathbb{P}(W_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = \mathbb{E}(W_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) =$$

$$= \mathbb{E}(\mathbb{E}(W_i \mid Y_i(0), Y_i(1), X_i, b(X_i)) \mid Y_i(0), Y_i(1), X_i, b(X_i)) =$$

$$= \mathbb{E}(\mathbb{E}(W_i \mid X_i, b(X_i))|Y_i(0), Y_i(1), X_i, b(X_i)) =$$

$$= \mathbb{E}(\mathbb{E}(W_i \mid b(X_i))|Y_i(1), X_i, b(X_i)) = \mathbb{E}(W_i \mid b(X_i)) = \mathbb{P}(W_i = 1 \mid b(X_i)),$$

where we used the law of total expectation, the assumption of unconfoundedness given the covariates and the definition of balancing scores.

■

As promised, the coarseness of the balancing score will also be shown, meaning that this is the best given the information contained in the covariates this is the best compression which preserves all the detail needed for balancing. To prove this we want to show that for any balancing score $b(x)$ there exists a function mapping it to the propensity score $p(x)$.

Proof. Coarseness of the propensity score.

We will prove the statement by contradiction. If there is no such function, then

$$\exists x_1, x_2 \text{ s.t. } b(x_1) = b(x_2) \text{ and } p(x_1) \neq p(x_2).$$

But then

$$\mathbb{P}(W_i = 1 \mid X_i = x_1) = p(x_1) \neq p(x_2) = \mathbb{P}(W_i = 1 \mid X_1 = x_2),$$

meaning that W_i and X_i are not independent given $b(x_1) = b(x_2)$, which contradicts the fact that $b(x)$ is a balancing score. ■

These characteristics of the propensity score guarantee that we can safely us it for reducing the dimensionality of the matching space while insuring the correctness and unbiasedness of our experiment. The only remaining issue is that in the case of relatively sparse data it is impossible to calculate the propensity score directly (if it would be possible we would already have sufficient amount of treatment and control pairs), so we need to estimate the score by some method. In the literature logistic regression, boosted CART and generalized boosted models are the ones most often used.[3]

After approximating the value of the propensity score we have successfully reduced the dimensionality of our balancing score, but the act of matching is still to be done. For these purposes the following methods are the most popular: nearest neighbors, optimal pair matching, ratio matching (not necessarily 1-to-1 and correcting for the inequality by weighting), subclassification (e.g. quintile grouping) and full matching (minimize distance within sets with at least one individual from the control and one from the treatment groups).[3]

The goodness of the matching is analyzed by comparing the distributions of the covariates (usually only marginal or pair-wise distribution). Another possibility is to check with regression-like methods if the treatment indicator can be predicted based on the covariates. The quality of the propensity score approximation is measured through evaluating the balancedness of the covariates in the resulting sample. Some authors consider using genetic algorithms for finding the best approximation method.[13]

This chapter took us through the theoretical preparation for our research and provides us with the needed concepts to starting dealing with the data and then applying these.

CHAPTER 2

DATA AND GENERAL STATISTICS

The data used for this research is a sample from a dataset already used for academic research before in the article entitled *Beautiful and damned. Combined effect of content quality and social ties on user engagement*.[2] In the first part of this chapter this work will be reviewed focusing on its the data-related parts and the results which are relevant to our approach as well. After presenting the process which resulted in the sample which is used throughout our research, some basic statistical properties will be revealed like distributions, trends, normalization, correlations and the time-based clustering of user activity.

2.1 Success and performance on a photo sharing platform

The data used in the research and the article being reviewed is from the online photo sharing-platform called Flickr, where users can upload many photos, tag them with meta-data, connect through friendships or follow others and interact with the uploaded content through comments and favorites. The platform released in 2004 has acquired tens of millions of users from which a sample of 40 million anonimized public profiles that are opted-in for research are considered with the 15 billion photos they have uploaded and more than half a billion social ties connecting them. The distributions of the main activity (number of photos uploaded) and popularity (favorite count) measures are presented in the original paper and shown here on figure 2.1.

The first very important output of the presented research is the creation of an aesthetic score which may serve as a performance metric besides the already available success or impact measures like favorites, views and comments. The process which yields this quality assessment is based on a deep neural network fine-tuned from a pretrained object detection network to classify images into buckets based on aesthetic considerations. Ground truth data for the aesthetic classification was collected through a crowd-sourcing approach ensuring independence from

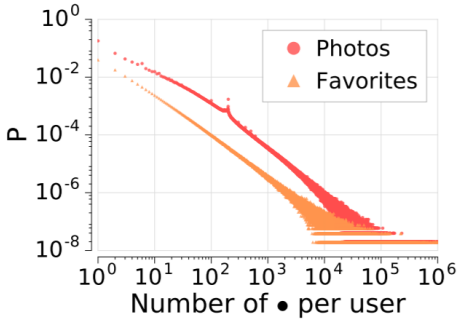

FIGURE 2.1. Figure showing the distributions of the main activity (number of photos uploaded) and popularity (favorite count) measures. Figure taken from.[11]

social relationship with the author and also concentrating on separation of personal impression from professional value. The output of the fine-tuned network is then converted into an aesthetic score through the following formula:

$$s = \frac{1}{2}(\mathbb{P}(\text{high}) - \mathbb{P}(\text{low}) + 1). \tag{2.1}$$

The second important part investigates interesting effects like correlation of quality and connectivity, majority-illusion effect (the illusion that the average user is at a lower quality level than the average of the friends, which is induced by local observations), assortative social ties based on user beauty (user beauty is defined as the average aesthetic scores of the images posted by the given user), new higher beauty connections causing average beauty increase and users having beauty scores far from the average of their neighbors leaving the platform with higher probability (churning problem).

The most interesting effect from our perspective is the causality related one about the user beauty enhancement caused by new connections. The authors of the article use a matching method where they iteratively remove the users which cause the highest differences between the covariate distributions until the predefined balance criterion (using the standardized bias as its central measure) is met:

$$SB_X(G_t, G_c) = \frac{\bar{X}_t - \bar{X}_c}{\sigma(X_t)} \leq 0.25, \tag{2.2}$$

where the specific value of the threshold is a commonly used one in the literature.[3]

After presenting the most relevant parts of research previously conducted on this dataset, the next step is to take an overview of the data tables which will be relevant for our purposes based on the questions we would like to answer. We also need to describe the exact meaning of the fields and the possible transformations needed to prepare them before we can go on from the data preprocessing and overview phase to the analysis of the phenomena behind them.

2.2 The sample and its structure

One of the main directions to be explored in this thesis is the use a propensity score based approach for imitating randomized experiments on data. For these purposes we will choose another setting focusing on the performance-success causality channel instead of the network-related effect considered in the article. For these purposes we will not use the network data, we will instead solely focus on the success measures and the aesthetic scores generated by the machine learning process described in section 1.2.

From the science of success perspective careers are the main objects of research. The careers have temporal dimensions, which requires multiple events happening in different points in time. To have data which makes it possible to talk about any kinds of statistics of these objects we chose to work only on the subsample containing users who have uploaded at least 10 photos. As the photo count has a broad distribution (figure 2.1) with many users uploading only very few pictures, this criterion has considerably reduced the data size (to approximately 100 thousand users and 85 million photos), making it possible to process on a personal computer (while still preserving the information which is valuable for us).

During the processing of the data further selection criteria were applied on top of this initial one: filtering based on upload date (based on information from the authors of the article considered in section 2.1,[2] only observations between 2006 and 2015 should be considered) and defining a unique upload date metric for setting the minimal activity limit for a career to be eligible for career analysis (details are described in section 2.4).

The initial format of the data was defined by the following two files containing the enumerated fields (in the parenthesis after the field name the variable type is also given). Data frame snapshots are shown in figure 2.2.

- photos.csv:
  - **pid** *(bigint)*: anonimized unique photo identifier
  - **date_imported** *(int)*: upload date of the photos in unix time, providing the time-embeddedness of the events and dynamics considered
  - **userid** *(bigint)*: anonimized unique user identifier
  - **count_comments** *(int)*: number of comments, measures impact through the number of verbal interactions or reactions evoked
  - **count_notes** *(int)*: number of notes, not necessarily a success measure, it is often added by the author of the photo to describe some details of the image
  - **count_tags** *(int)*: number of tags, not necessarily a success measure, it is usually added by the author of the photo as metadata (can discribe objects, places, events, etc.)
  - **count_faves** *(int)*: number of favorites, measures success through the number of positive reactions evoked

(a)

	pid	date_imported	userid	count_comments	count_notes	count_tags	count_faves	count_views
0	[redacted]	2015-06-18 11:22:08	[redacted]	0	0	8	7	269
1		2015-06-18 11:52:38		8	0	16	29	1040
2		2015-06-18 12:20:23		0	0	0	1	23
3		2015-06-18 12:20:23		1	0	0	3	23
4		2015-06-18 12:20:23		0	0	0	1	20
5		2015-06-18 12:20:23		0	0	0	0	14
6		2015-06-18 12:20:24		0	0	0	2	15
7		2015-06-18 12:20:24		0	0	0	0	8
8		2015-06-18 12:20:25		0	0	0	0	10
9		2015-06-18 12:20:25		0	0	0	0	8

(b)

	pid	score	userid
0		0.106	
1		0.085	
2		0.155	
3		0.245	
4		0.206	

FIGURE 2.2. Snapshots of the pandas data frames created from the photos.csv (a) and aesthetic_scores.csv (b) files. (The unique identifiers are pixelated to ensure maximal privacy.)

  - **count_views** *(int)*: number of views, measures success and visibility

- aesthetic_scores.csv:
  - **pid** *(bigint)*: anonimized unique photo identifier
  - **score** *(float)*: aesthetic score value, being the performance (quality) metric used in our research to compare to the success measures defined in the photos.csv file
  - **userid** *(bigint)*: anonimized unique user identifier

Might be important to clarify that there is no temporal dimension available for the views and interactions of the photos in the current data, we only see the cumulative values of these scores until the time (March 2016) when the observations were queried and provided as a dataset.

Since we are interested in comparing success and quality of the photos, the unique photo identifier will be used to match the two datasets.

Before being ready for this step some basic sanity checks and preprocessing was applied to eliminate small data errors like invalid rows, duplicated identifiers or non-standard encoding of some fields.

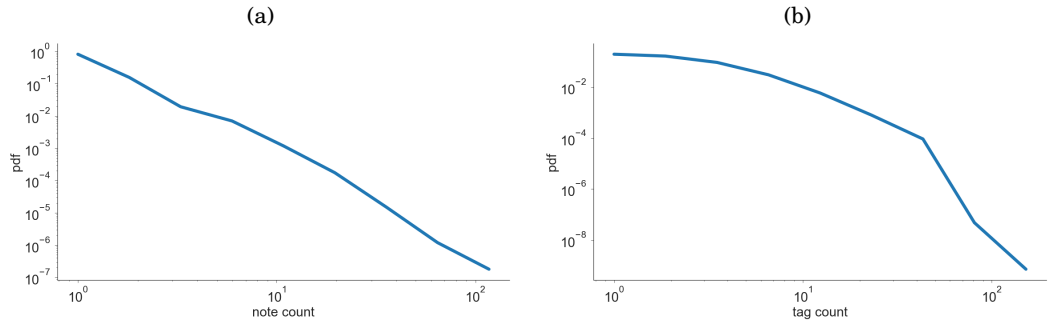

FIGURE 2.3. Note count (a) and tag count (b) distributions on log-log scales.

When working with data one should always fit the computational cost of the investigations that he wants to perform with the available resources. In the actual case the computations were done on a personal computer, a fact which sets stricter limitations on memory size and computational power. Thus having around 85 million photos makes it impossible to join the two tables directly, which would mean an $O(n^2)$ complexity. Rather some kind of divide and conquer approach is needed: if one first sorts both of them by the unique photo identifier in $O(n \log(n))$, the only $O(n)$ is required for the merge.

2.3 Data field distributions

Firstly, the distributions of note count (2.3(a)) and tag count (2.3(b)) are presented. Here we can observe that although the distributions are close to linear when presented in a log-log setting, the maximal values do not exceed 100 significantly. This observation can be easily understood considering that these measures refer to the amount of metadata provided by the author in most of the cases (other users might add notes and tags as well if the permissions are set up in such a way).

As a next step, the distributions of aesthetic score (2.4(a)), comment count (2.4(b)), view count (2.4(c)) and favorite count (2.4(d)) are shown. These four aligned plots allow us to identify the contrasting nature of performance (aesthetic score) and success/impact/visibility (comments, views, favorites) measures. The popularity measures all have very broad distributions, spreading across orders of magnitude as a consequence of the multiplicative nature of the underlying dynamics. The aesthetic score on the other hand is a measure of quality, which was already defined to be well-bounded (between 0 and 1), but if we take a look at the scale of the density values we can see that the distribution is not that skewed, there is roughly just a 10-times factor between the densities at the two ends of the value range.

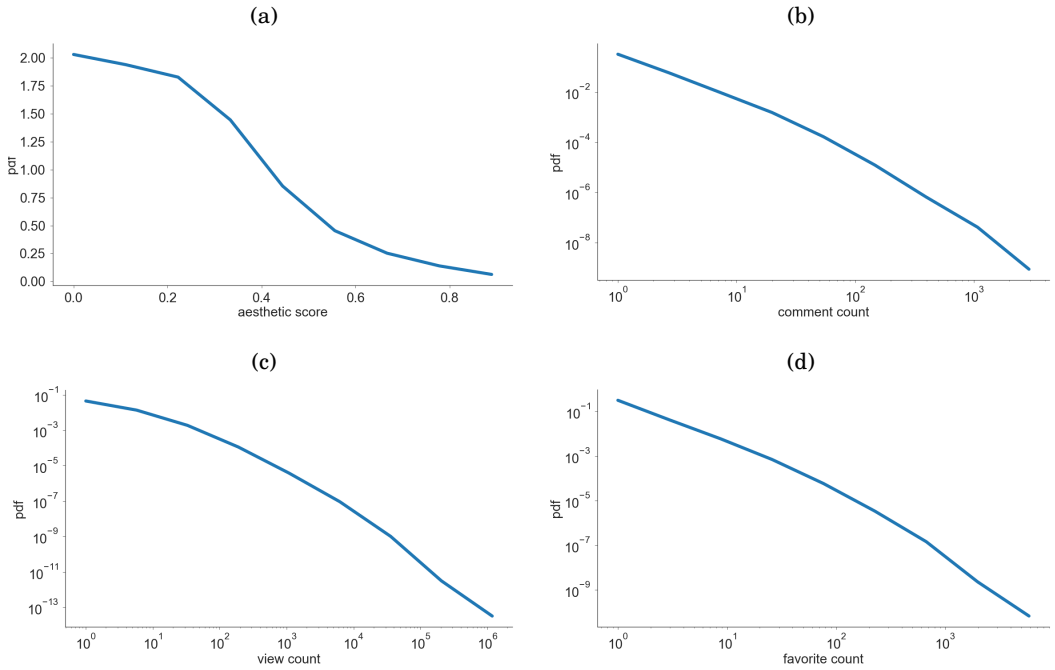

FIGURE 2.4. Distribution of the performance measure (aesthetic score – a). Comment count (b), view count (c) and favorite count (d) distributions on log-log scales.

2.4 Unique upload dates and defining albums

Quite early in the research process we realized that users tend to upload photos in batches and such batches can have quite a lot of photos in them. Thus our first assumption that 10 data-points could already be considered a career does not hold, as even 100 photos may be uploaded in a singe batch, but we still cannot talk about any kind of trends or career paths until having a reasonable amount of upload sessions.

So we defined the concept of upload dates, which simply refers to the days when the user uploaded a positive number of photos, and required that the careers considered have at least 10 such events. Even this way there might be careers with very different time scales, but at least we can be sure that the remaining data consists of events which happened at different points in time.

After observing that the photos are often uploaded in batches or at least quite close to each other in time, we might argue that in fact such a batch should be considered one piece of work. A photographer may go to a location and take several hundreds of photos a day and upload the best ones at the end of the day for example. On the other hand it might happen that some photos are uploaded in one batch or very close in time but still differ in metadata like location, theme or style.

To better understand the situation regarding the upload times, figure 2.5 shows the distribu-

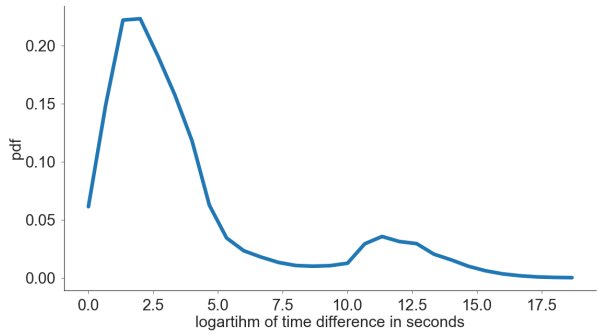

FIGURE 2.5. The distribution of time passed between two consecutive photo uploads by the same user. Time is measured in seconds and the natural logarithm of this value is displayed.

tion of time differences, e.g. the time that passed between two consecutive photo uploads from the same user.

As it is clear from the plot, there is a low-density regime for time differences at $5 \leq \ln \Delta t \leq 10$ (time being expressed in seconds). This might be interpreted as a separation between the photos uploaded in the same session or in different ones. Considering the albums as being the basic building blocks of a career rather then the photos themselves could be a viable alternative through defining the borders of these "albums" to be where time differences of more than 1.5 hours ($\ln \Delta t^*(s) \approx 8.5 \iff \Delta t^* \approx 1.5h$) appear between consecutive uploads.

2.5 Trends and normalization

The next issue which needed to be addressed was the elimination of the overall detectable trends in the data. The following figures illustrate the significance of these trends and the average data range for some important measures: number of photos uploaded in 100 days (figure 2.6(a)), mean aesthetic score (figure 2.6(b)), mean view count (figure 2.6(c)) and mean favorite count (figure 2.6(d)). The averages of the measures considered might vary over time depending on the activity patterns and habits of the user-base. We would like to exclude these effects so we can examine the phenomena which would be present in a stationary state of the system independently of these factors. Note that given the way these measures are available to us (cumulative until the point when the data was acquired) a decreasing trend would be expected in a stationary system (the old photos have more time to acquire views, comments, favorites, etc.). The trends observed in our case are mostly upward ones thus the system must go through a strongly non-stationary period.

To make sure these effects do not distort our results we have applied normalization: division by the mean,[14] where the average was calculated for a window centered around the observation to be normalized. These rolling calculations reduce the possibility of numerical artifacts impacting our findings.

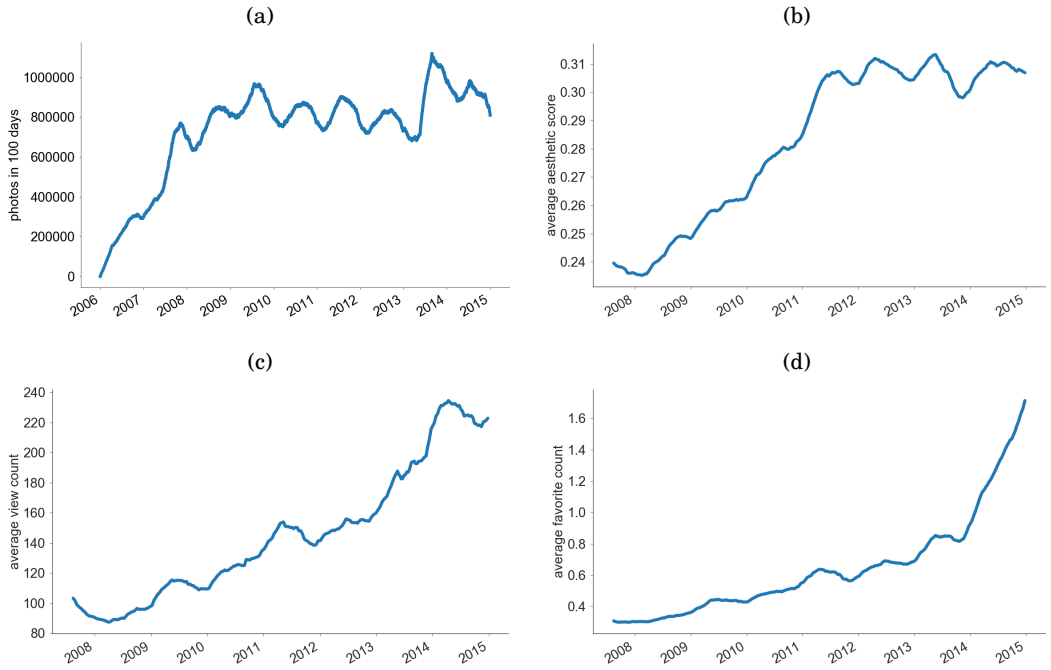

FIGURE 2.6. Number of photos uploaded in 100 days (a – rolling trend). Average aesthetic score (b), view count (c) and favorite count (d) of 2 million uploaded photos (rolling trend).

2.6 Correlations

As usually in the case of correlations calculated on data with broad distributions we opt for calculating the Spearman correlation coefficient, which means calculating the Pearson-correlation of the ranks of the input values. The following formula describes the procedure in mathematical notation:

$$r_{Spearman}(X,Y) = \rho(\bar{X},\bar{Y}) = \frac{cov(\bar{X},\bar{Y})}{\sigma(\bar{X})\cdot\sigma(\bar{Y})}, \tag{2.3}$$

where $\bar{X}$ means the ranked version of the variable X, e.g. mapping the values to a uniform distribution on $(0,1)$ through sorting.

We calculate correlations between our fields in two different ways: first on a photo level directly (figure 2.7(a)) then after aggregating the values on user level (the mean of the career, shown in figure 2.7(b)). The aggregation procedure makes the correlations more pronounced, an expected result of averaging through noise reduction.

We can see from the plot that the main success measures: comment count, view count and favorite count correlate between themselves quite strongly. As expected, note count and tag count are quite different form these three. Our performance measure, the aesthetic score is in strong correlation with the favorite and comment count measures, but it is less coupled with view count

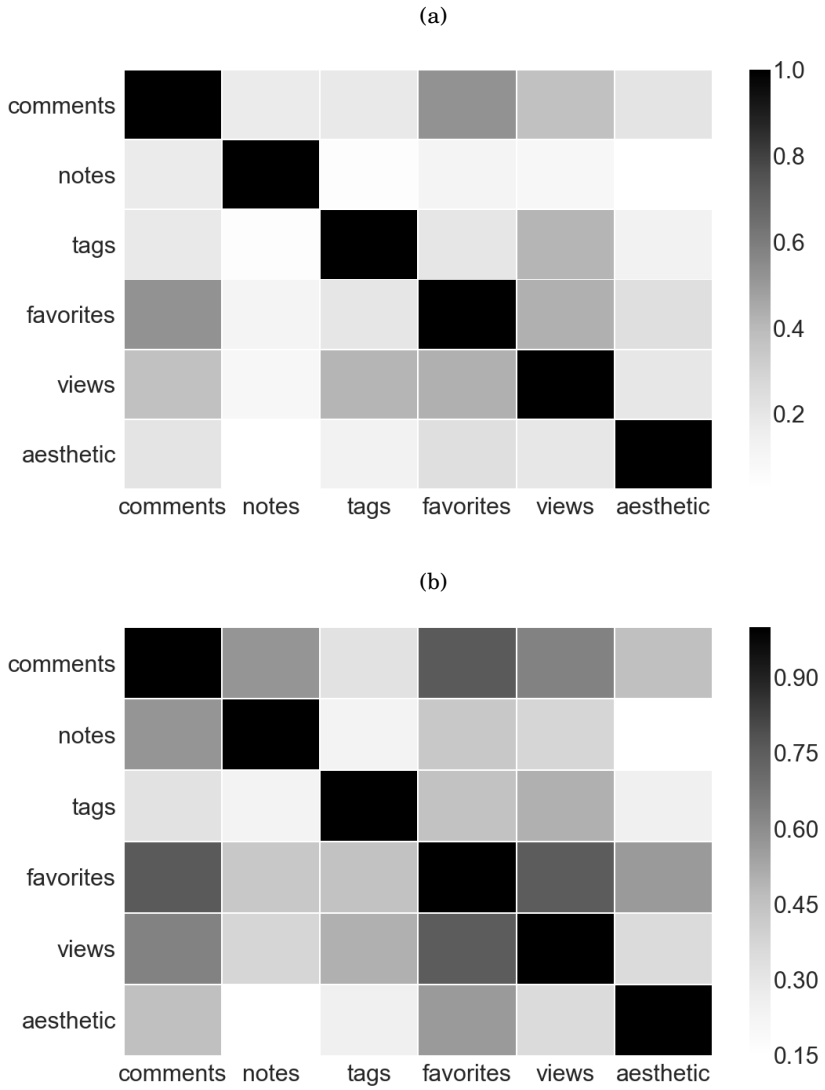

FIGURE 2.7. Correlation heatmap of the considered metrics – comment count, note count, tag count, favorite count, view count and aesthetic score – on photo (a) and user aggregated (b) level.

(also understandable, as view count is more or less stable throughout a career period, while the aesthetic value of the uploaded photos and the interaction generated might vary a lot even within a single album).

CHAPTER 3

RESULTS

This chapter will present the results of the data analysis in comparison with the results found in the literature and discuss the differences and similarities. Specifically it aims at presenting the details of two experiments: first calculating the random impact curves, comparing them to null models and trying to explain the found differences; and secondly the application of the propensity score based approach to infer the presence of causality between performance and success.

3.1 Random impact rule

To start with, the random impact rule, as presented in section 1.1, is the statement that the timing (more exactly ranking) of the most successful hit during a career is random (uniformly distributed). In mathematical notation that is:

$$\frac{N^*}{N} \sim U(0,1), \tag{3.1}$$

where N^* is the temporal rank of the best (highest in some success measure) product, N is the total number of products during the career and U denotes the uniform distribution.

The basis of comparison for the results is given by a null model, which in this context is obtained by reshuffling the values of the measure selected for analysis. More specifically we preserve the timestamps of the events (uploads), but we reshuffle the performance or success measures associated with the photos. The reshuffling can be done either within the careers of each user or on the entire dataset (both were tried and no significant difference was observed in the case of our analysis; within-career reshuffling is used in the presented results).

An interesting situation was encountered during the research process: in the case of favorite count many similar values were found (discrete variable with typically low values), and it often

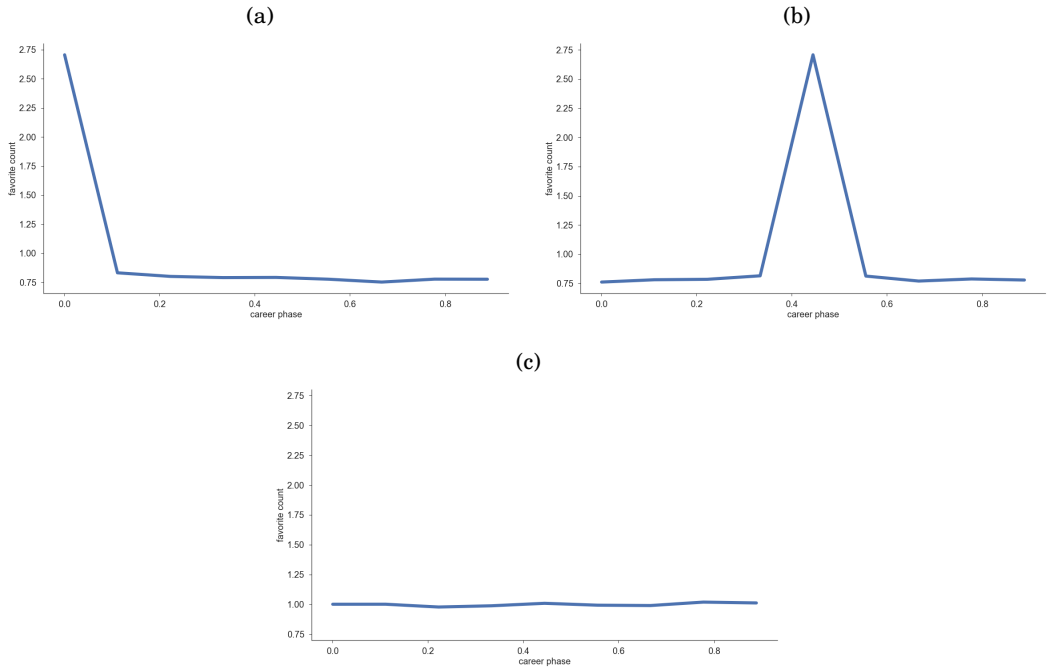

FIGURE 3.1. Distribution of N^*/N in the case of favorites count after reshuffling the values within careers. First the two biased versions are shown: always selecting the first maximum (a) and considering the mean of the positions of the maxima as being the position of the maximum (b). Finally the unbiased version is presented: selecting the representative for the summary at random (c).

happened that the highest values coincided as well. In such situations the distribution was totally distorted by numerical artifacts even in the randomized case. Figure 3.1 is presenting the process of searching for the right solution to this issue.

In the beginning simply the first of the maxima was considered naively, which resulted in getting a higher density in the beginning of the career (figure 3.1(a)).

As a not too inspired trial towards the solution was to attribute the position of the maximum to the mean of the positions with maximal values, but this led to the apparition of a peak in the middle of the distribution (as expected, when looking at in retrospective, shown in figure 3.1(b)). So finally the good solution was to select the position of the maximum randomly from the available maxima (result shown in figure 3.1(c)), as clearly we have to choose uniformly from a sample if we would like to obtain an un-biased estimate of the original distribution after accumulating the results of several such trials.

After solving these numerical issues the comparison of the results obtained on the original data and the null model is ahead. This was done by splitting the entire dataset into 10 subsamples and calculating the values for each bin (career period) based on all these subsets. In the end the plot shows the averages of the measurements surrounded by the shaded areas associated with

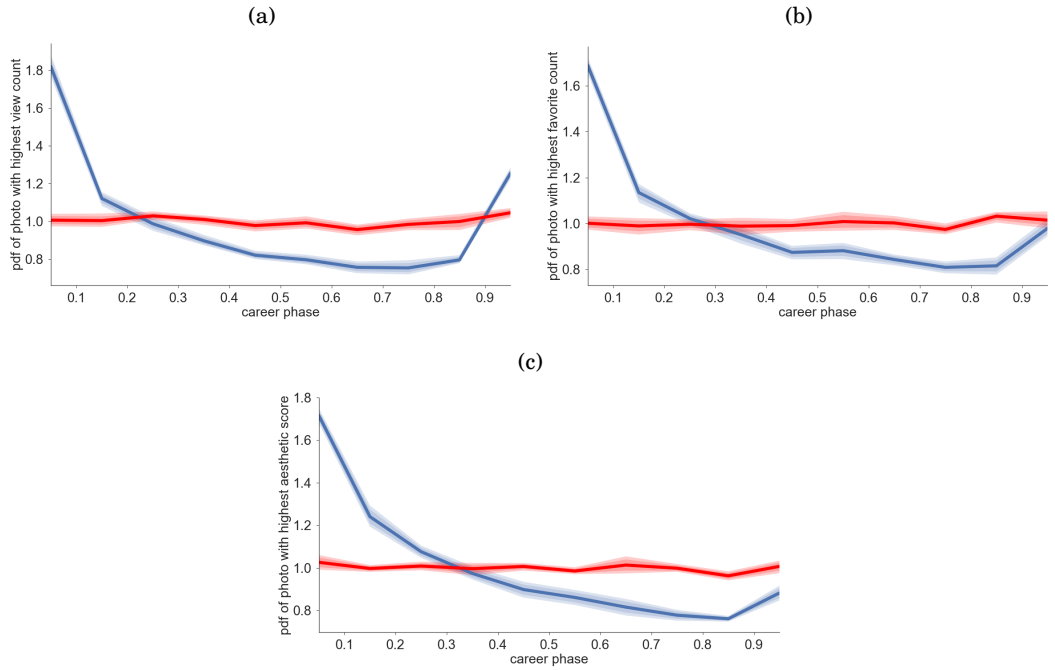

FIGURE 3.2. Distribution of N^*/N in the case of the most important success measures – view count (a), favorite count (b) – and the performance measure – aesthetic score (c).

the 90% and 99% significance levels, which were estimated through bootstrapping.

The measurement was done for the most important success measures – view count (figure 3.2(a)), favorite count (figure 3.2(b)) – and the performance measure – aesthetic score (figure 3.2(c)).

The plots clearly show that the distributions aren't identical with those from the null model, meaning that for some reason there is higher probability that a user will upload his most successful (and also the most aesthetic) photo in the beginning of the career and there is less chance that this will happen after the middle and before the end of the career.

3.2 Possible causes, hypothesis and analysis

As this finding is not in alignment with those resulting from the study of scientific careers, it is natural to hypothesize that there is some fundamental difference between the two fields. Indeed, we already observed in the data analysis part that the photos are often uploaded in batches. Maybe the structure of the batches changes during the career. It might as well happen that there are different types of users (e.g. which are most popular in the beginning of their career opposed to those most popular in the later stage). Or perhaps there is some phenomenon which is not deductible from the data but might cause such deviations. The following sub-sections will

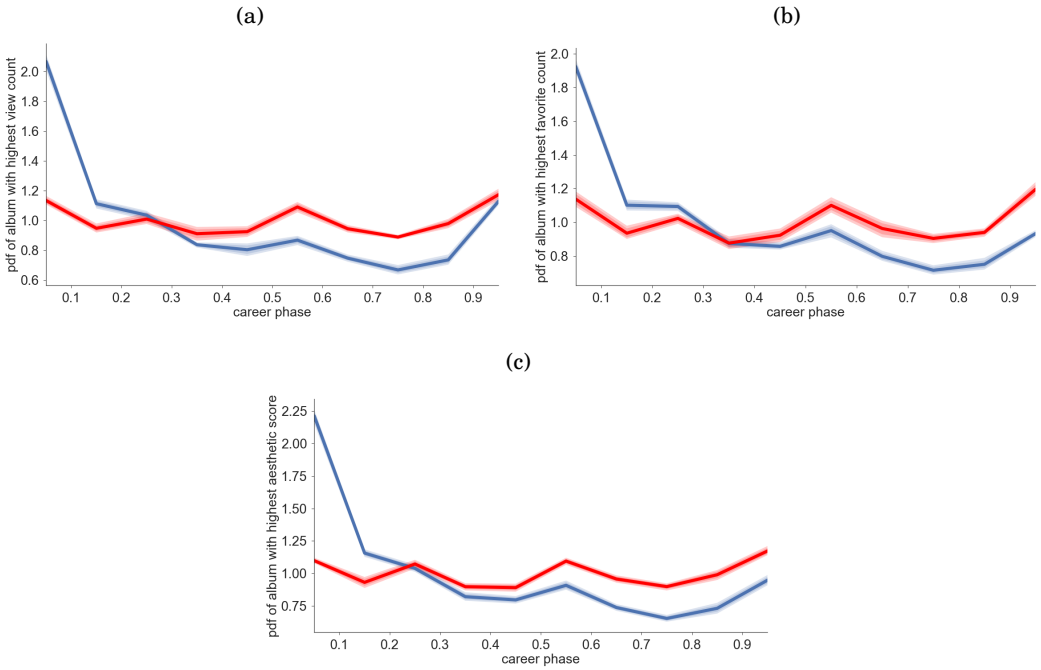

FIGURE 3.3. Distribution of N^*/N in the case of the most important success measures – view count (a), favorite count (b) – and the performance measure – aesthetic score (c) – after grouping the photos into albums and associating with the groups the maximal value attained for each of the measures.

concentrate on exploring (or starting to explore) some of these directions.

3.2.1 Album-based analysis

Based on the first observations regarding the time-wise clustering of user activity, which is a clear difference from the usual scientific productivity patterns, one could argue that the experiment should be done by considering an album to be the elementary building block of a career instead of a single photo. The main limitation of this approach is that we do not have direct data about the albums defined by the users (and even if this data was available, not all users use the possibility to group their work in albums).

But for a first order approximation it is possible to do the album classification based on the approximately 1.5 hour delta in the upload times and see what happens to the previous results. The findings are presented in figure 3.3.

As seen from the plots this album based approach did not help in finding out the reason behind the divergent behavior.

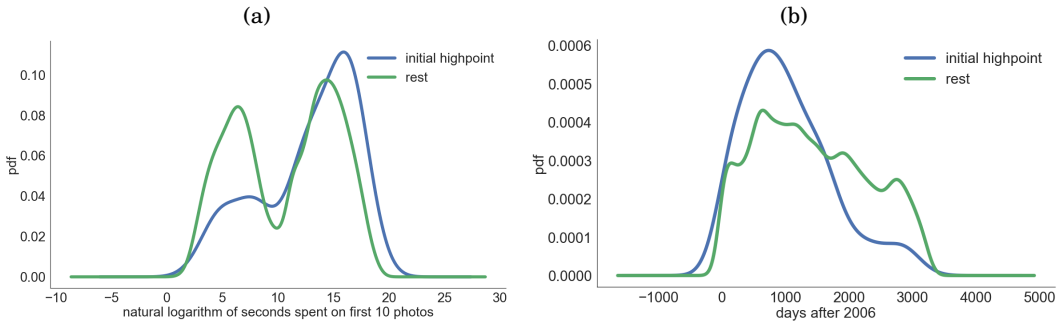

FIGURE 3.4. Comparing the characteristics of users who had their highest hit in the first 5% of their career based on aesthetic score, view count and favorite count as well. The plots show the time it took them to upload their first 10 photos (a) and when did they upload their first photo (b). Note: kernel density estimation was used, this may cause the distributions to stretch into the negative region.

3.2.2 User profiling

The second approach which may help in finding the cause of the non-usual career high point distribution would be to analyze if there is some significant difference in the distributions of parameters of the users which have their success in the very beginning of the career compared to the rest.

Firstly, the users which had the highest score in the first 5% of their career were selected based on three different scores: aesthetic score, view count and favorite count. For all of these there were roughly 2 times as many users falling into this category as if the distribution were to be uniform. After this the first obvious question to ask was if the overlap of these sets is significantly higher than random. It turns out that the intersection of the 3 sets is 5 times as big as it would be if the members would be selected at random.

Now, to see if there are some common characteristics of these users compared to the rest of the population, as an example the length of the time interval in which the first 10 photos were submitted (figure 3.4(a)) and the beginning of the career (days since 2006 – figure 3.4(b)) were selected as reference measures.

As illustrated by the plots, the users from the analyzed group upload their first 10 photos during a longer time period (twice as long on average) and tend to have started their career earlier in the history of Flickr than the average user (400 days earlier on average). Taking into account these considerations we may hypothesize that the longer time spent during the first 10 photos means that it is not a single big batch, but maybe a diverse selection of photos or that the observed phenomenon is somehow connected to the early period of the platform. This kind of analysis could help in coming up with further research directions which could later be explored and possibly lead to some answers to the questions asked.

These are just a few examples of trying to dig deeper into this phenomenon, but similar

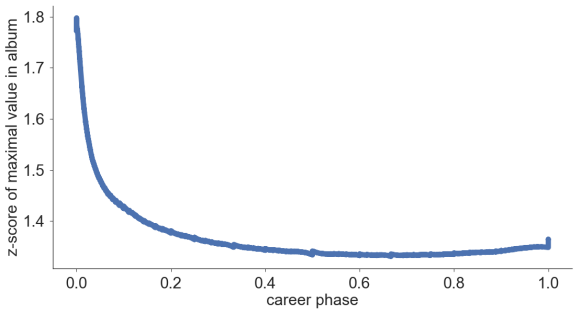

FIGURE 3.5. Average z-score of the maximal elements from each album with respect to career phase.

analysis could be used in general when trying to understand the reasons behind deviations from the expected behavior.

3.2.3 Diversity

The third suggested approach is trying to follow the gut feeling that users might submit selections of older photos as well in the beginning of their career. This will boost the probability that one of these uploads would be the most successful in the entire career (simply because those represent a longer period with more photos taken in reality but not uploaded to the platform). This hypothesis is a very natural one, but very hard to capture in the currently available data. As a first attempt to localize a footprint of the hypothesized behavior outliers might be of some use. If it is true, that usually albums contain photos taken in similar circumstances except for the first ones, the beginning of the careers should contain more photos which are outliers or at least very different from the rest of the album (batch of uploaded photos). As a proxy towards capturing this diversity, the photo having the highest aesthetic score is considered from each album and it's z-score is calculated with respect to this group. The average z-score of the maximal elements is shown along the career (calculated through exponential moving average – figure 3.5).

The results from this approach tend to underline our belief about the different upload behavior at the initial stage of the career, which in fact would distill the work of a much longer period with a lot more photos into a selected group of a few uploaded ones.

3.3 Causality analysis

In this section the aim is to show a simple example of applying the method of matching based on propensity score to a scenario involving the relationship between performance and success. As one of the most appealing features of the used dataset is that it contains both measures of photo

quality and popularity, it's a straightforward idea to try proving a causal relationship between these different kinds of measures.

We can think of the aesthetic score associated to a photo as an objective measure related to the probability that a random observer would enjoy watching this image (based on the ground truth data used for training the network). As such, maybe the simplest experiment would ask the following question: does higher than average aesthetic quality imply in a causal manner the higher probability of positive interaction from a user (to proxy this probability the ratio between favorite count and view count of a photo is used, i.e. how many of the users who have seen it have considered it beautiful enough to mark it as favorite).

As seen in section 1.3, besides defining the treatment and effect variables, defining covariate variables to capture the circumstances which could interfere with the treatment or its effect is crucial. To exclude as many other causal relationships as possible – for example initial success may have severe consequences on the future career path or network embedding could determine the evolution of the follower base – the beginnings of the careers were considered, when there is no history that could affect our experiment. For these reasons the first 10 photos of each user are used for this analysis and based on these the following covariate factors are defined: the time interval stretched by these 10 photos, upload dates (days measured from 2006, when our data begins), in how many "album" sessions were they uploaded (see section 2.4), what is the average tag and view count respectively.

The next step is to approximate the propensity score. For the sake of simplicity linear regression was chosen for this purpose. The target variable is the treatment (the average aesthetic score of the first 10 photos being above or below average) expressed as 0 or 1. To enhance the possibilities of the model towards capturing the relationships between the data-points we reduce the heterogeneity of the dependent variables by transforming the view count and favorites count into logarithmic space before using them in the regression. Even after this small help, the fit isn't impressive ($R^2 = 0.03$), but still the such approximated propensity score proves to be useful at balancing the covariate distributions: bins of size 0.05 in predicted propensity score are considered and equal number of treated and control cases are added to the analysis from any such bin. More explicitly, if $n_t(b)$ and $n_c(b)$ denote the number of treatment and control individuals in the given bin b, then

$$n(b) = \min(n_t(b), n_c(b)) \tag{3.2}$$

members are added from both groups (from the larger group we select the individuals which are discarded at random). This way the control and treatment will have the same size and it is easier to apply statistical significance tests in the end and do not have to use frequency weighting on them. To see the effects of balancing the probability densities of the following measures are shown in the treatment and control groups before and after the balancing: the approximated propensity score (figures 3.6(a) and 3.6(b)), the album count (figures 3.6(c) and 3.6(d)) and the

	SB before	SB after
propensity score	0.357	0.016
average view count	0.254	0.045
album count	0.241	0.014
average tag count	0.205	0.044
days since 2016	0.094	0.038
time interval	0.049	0.006

Table 3.1: Table showing standardized bias of the covariates before and after balancing.

view count(figures 3.6(e) and 3.6(f)). The results can be expressed in terms of the reduction achieved in standardized bias (2.2) in each case, shown in table 3.1.

Remarkably, to obtain these results, only one fifth of the data was discarded, one of the great advantages of using the propensity score, the coarsest balancing score of all. (We could have discard even less data if we would not persist on having the same number of individuals for each bin of the propensity score).

And finally, collecting the fruits of our balancing work, the results of the experiment can be evaluated. First, to show the resulting positive interaction (favorites count per view count), the distributions are plotted (figures 3.7).

Although it is clearly visible form the plot that the treatment causes higher probability of positive interaction, in fact the mean positive interaction probability is 3 times as high for the treatment group than for the control, the statistical significance of this claim has to be addressed. For this purpose the Wilcoxon-Mann-Whitney test is applied, which is a nonparametric test suitable for assessing the null hypothesis that the two samples come from the same distribution.[15] Applying this test to our treatment and control data, the null hypothesis is refuted and the alternative hypothesis of the treatment data containing higher values than the control data is accepted with significance corresponding to $p < 0.00001$ (sample size $n_1 = n_2 = 17978$ with rank-sum of $R = 195180952.5$ which would correspond to a z-score in the order of magnitude of tens of thousands).

Thus we can conclude that our causality analysis proved the implication between aesthetic value of an uploaded photo and the willingness of positive interaction by the users viewing it. Clearly, this was just a simple example of applying this methodology on performance-success dynamics, but it already shows the power of the approach and it serves as a proof of concept to encourage further similar applications.

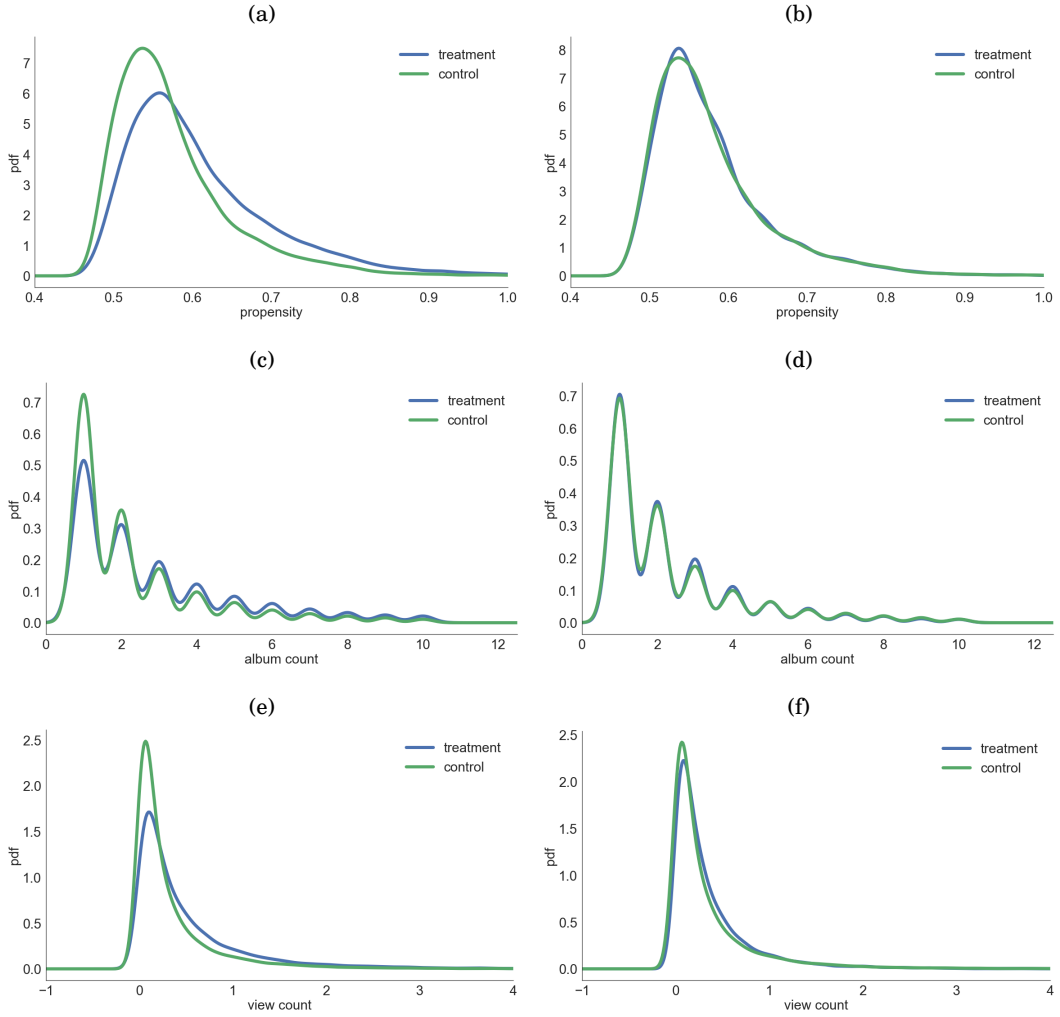

FIGURE 3.6. Propensity score distribution originally (a) and after balancing (b). Album count distribution originally (c) and after balancing (d). View count distribution originally (e) and after balancing (f). Note: kernel density estimation was used, this may cause the distributions to stretch into the negative region.

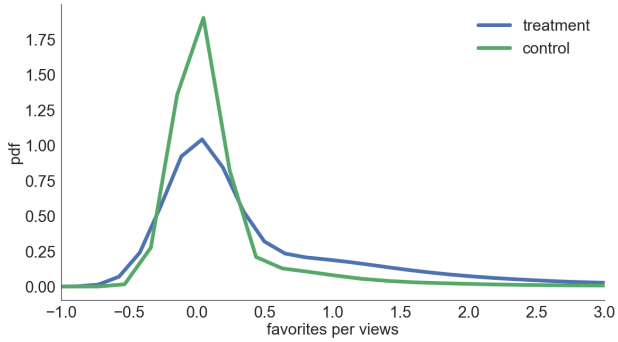

FIGURE 3.7. The distribution of positive interaction (favorites count per view count) in the control and treatment groups. Note: kernel density estimation was used, this may cause the distributions to stretch into the negative region.

Conclusion

To achieve the first goal, analyzing success dynamics of online photos sharing, the methodology related to career analysis and the random impact rule was used, the distributions of the different success measures were plotted, finding that they indeed span across multiple orders of magnitude, as expected based on the multiplicative nature of the underlying dynamics. One very important difference was found compared to previous results from the field of science of success: in the case of photographer's careers the distribution of maximal success (and performance) is not uniform (at least by direct computation).

Several possible causes of the fact that there is higher than average probability of top success in the beginning of the career were described. The photos are usually uploaded in batches, which made it possible to define albums based on upload sessions (upload date clustering), which might imply deviations between the results, but this direction did not bring us closer to understanding the phenomenon.

The users which have their biggest hit in the very beginning of their career have significantly different characteristics from the rest: they usually upload their first 10 photos during a longer time period and start their career earlier. This kind of analysis can give more insight about the possible causes of the anomaly, thus helping in developing further hypothesis.

Lastly, the most trivial hypothesis is that in the beginning of the career photographers upload selected photos from their previous works, which this way concentrate a lot of value in a small amount of pictures, thus raising diversity and the probability of success. This kind of assumption is very hard to assess based on our data, but a possible direction was pointed out, which based on the initial results supports this claim.

Finally, the second goal of the research was to deduct randomized experiments on the observational data available to assess the causal relationship between performance and success. In particular the question whether appealing aesthetic quality raises the probability of positive interaction was asked. Using matching through propensity score the covariate distributions of the treatment and control groups were successfully balanced (the standardized bias was reduced both below the often used 0.25 and 0.1 thresholds) and the experiment confirmed the expected causal tie between aesthetic score and the average number of favorites given per view.

Thus in the end we might conclude that the research attained its goals, managed to use knowledge coming from several fields to better understand the data and the phenomena behind it. It yielded interesting results as well as findings which might encourage further research and

the adoption of the adequate mathematical tools for inferring causality based on observational data.

APPENDIX A

APPENDIX

The appendix is meant to present some code snippets which define important measures, create some of the fundamental plots or reveal additional technical details about the research process. First of all, the research was done in Python and the most widely used packages were the following: pandas, numpy, scipy, matplotlib and seaborn.

The first thing which was not directly solved by the already available functionality was to set a threshold on the minimum number of days in which each user was active:

```
unique_dates = photo_data[['userid','date_imported']].groupby('userid')
                    .apply(lambda x:
        len(set([y.strftime('%Y%m%d') for y in list(x['date_imported'])])))
```

Then this measure was used to filter for the users having at least 10 unique upload dates.

To join the tables of performance and success measures, first both of them were sorted based on pid (unique photo id), then the following procedure was used to select the intersection in linear time:

```
aesth_i = list(aesth_data_i.index)
photo_i = list(photo_data_i.index)
intersect = []
i = 0
j = 0
while i < len(aesth_i) and j < len(photo_i):
    if aesth_i[i] == photo_i[j]:
        intersect.append(aesth_i[i])
        i += 1
        j += 1
```

```
    elif aesth_i[i] > photo_i[j]:
        j += 1
    else:
        i += 1
```

For the trends and normalization part basic rolling window features were used from the pandas package. The plotting functionality was defined in a customized way, being able to use already included plotting functions from pandas and seaborn (like plot, hist, plot.kde, etc):

```
def plot_processor (d, path, xlabel, ylabel, fontsize = 30, figsize=(16,9),
                    style = 'white', **kwargs):
    fig = plt.figure(figsize=figsize)
    ax = fig.add_subplot(111)
    d(ax=ax, **kwargs)
    ax.set_ylabel(ylabel, fontsize=fontsize)
    ax.set_xlabel(xlabel, fontsize=fontsize)
    ax.tick_params(labelsize = fontsize, pad = 10)
    sb.set_style(style)
    sb.despine()
    plt.tight_layout()
    plt.savefig(path)
    plt.show()
```

The calculation of the N^*/N distribution for the random impact rule applied the following function on a groupby for the users:

```
def calc_random_impact(frame):
    return frame.groupby(level=0).apply(lambda x: 1.0*
            (np.random.choice(
                np.argwhere(
                    np.array(x.values == np.amax(np.array(x.values)))
                ).flatten()
            ))/(len(x)-1))
```

After defining the values which needed for the assessment of the distributions, the following function was used for bootstrapping and plottting:

```
def ri_plotter(**kwargs):
    rs = np.split(r.sample(frac = 1.0).values[:-(len(r)%m)], m)
    r2s = np.split(r2.sample(frac = 1.0).values[:-(len(r2)%m)], m)
    ds = []
    d2s = []
    for i in range(m):
```

```
        d, _ = np.histogram(rs[i], bins = binc, density = True)
        ds.append(d)
        d2, _ = np.histogram(r2s[i], bins = binc, density = True)
        d2s.append(d2)
    sb.tsplot(np.array(ds), time = [(i+0.5)/binc for i in range(binc)],
              ci = [90,99], **kwargs)
    sb.tsplot(np.array(d2s), time = [(i+0.5)/binc for i in range(binc)],
              ci = [90,99], color = 'red', **kwargs)
```

Finally, the causality related investigations used scipy functionality for fitting the linear regression and assessing statistical significance, while the following function served to prepare the graphs:

```
def treatment_control_density(tr_list, co_list, ax, df, **kwargs):
    tr = df[df.index.isin(tr_list)]
    co = df[df.index.isin(co_list)]
    tr.plot.kde(label = 'treatment', ax = ax, **kwargs)
    co.plot.kde(label = 'control', ax = ax, **kwargs)
    print (abs(tr.mean()-co.mean())/tr.std())
```

The above code snippets are meant to give enough additional detail about the work and a taste of the programming work involved. If any further questions would come up, please feel free to contact the author.

BIBLIOGRAPHY

[1] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási.
Quantifying the evolution of individual scientific impact.
Science, 354(6312):aaf5239, 2016.

[2] Luca Maria Aiello, Rossano Schifanella, Miriam Redi, Stacey Svetlichnaya, Frank Liu, and Simon Osindero.
Beautiful and damned. combined effect of content quality and social ties on user engagement.
IEEE Transactions on Knowledge and Data Engineering, 29(12):2682–2695, 2017.

[3] Elizabeth A Stuart.
Matching methods for causal inference: A review and a look forward.
Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1, 2010.

[4] Burcu Yucesoy and Albert-László Barabási.
Untangling performance from success.
EPJ Data Science, 5(1):17, 2016.

[5] Sherwin Rosen.
The economics of superstars.
The American economic review, 71(5):845–858, 1981.

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási.
Internet: Diameter of the world-wide web.
nature, 401(6749):130, 1999.

[7] Albert-László Barabási and Réka Albert.
Emergence of scaling in random networks.
science, 286(5439):509–512, 1999.

[8] Mark EJ Newman.
Power laws, pareto distributions and zipf's law.
Contemporary physics, 46(5):323–351, 2005.

[9] Deanna Blansky, Christina Kavanaugh, Cara Boothroyd, Brianna Benson, Julie Gallagher, John Endress, and Hiroki Sayama.

Spread of academic success in a high school social network.
PloS one, 8(2):e55944, 2013.

[10] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang.
Studying aesthetics in photographic images using a computational approach.
In *European Conference on Computer Vision*, pages 288–301. Springer, 2006.

[11] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang.
Rapid: Rating pictorial aesthetics using deep learning.
In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014.

[12] Guido W Imbens and Donald B Rubin.
Causal inference in statistics, social, and biomedical sciences.
Cambridge University Press, 2015.

[13] Alexis Diamond and Jasjeet S Sekhon.
Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.
Review of Economics and Statistics, 95(3):932–945, 2013.

[14] Filippo Radicchi, Santo Fortunato, and Claudio Castellano.
Universality of citation distributions: Toward an objective measure of scientific impact.
Proceedings of the National Academy of Sciences, 105(45):17268–17272, 2008.

[15] Henry B Mann and Donald R Whitney.
On a test of whether one of two random variables is stochastically larger than the other.
The annals of mathematical statistics, pages 50–60, 1947.