Beautiful and damned. Combined effect of content quality and social ties on user engagement

This is a pre print version of the following article:

*Original Citation:*
Beautiful and damned. Combined effect of content quality and social ties on user engagement / Aiello, Luca Maria*; Schifanella, Rossano; Redi, Miriam; Svetlichnaya, Stacey; Liu, Frank; Osindero, Simon. - In: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. - ISSN 1041-4347. - 29:12(2017), pp. 2682-2695.

*Availability:*
This version is available http://hdl.handle.net/2318/1662640 since 2018-03-17T10:07:57Z

*Published version:*
DOI:10.1109/TKDE.2017.2747552

*Terms of use:*
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.
Beautiful and damned. Combined effect of content quality and social ties on user engagement

Luca M. Aiello
Nokia Bell Labs
luca.aiello@nokia-bell-labs.com

Rossano Schifanella
University of Turin
schifane@di.unito.it

Miriam Redi
Nokia Bell Labs
miriam.redi@nokia-bell-labs.com

Stacey Svetlichnaya
Flickr
stacey@yahoo-inc.com

Frank Liu
Flickr
frank@yahoo-inc.com

Simon Osindero
Flickr
simon@yahoo-inc.com

Published in IEEE Transactions on Knowledge and Data Engineering (Volume: PP, Issue: 99). Available at https://doi.org/10.1109/TKDE.2017.2747552.

ABSTRACT

User participation in online communities is driven by the intertwine-ment of the social network structure with the crowd-generated content that flows along its links. These aspects are rarely explored jointly and at scale. By looking at how users generate and access pictures of varying beauty on Flickr, we investigate how the production of quality impacts the dynamics of online social systems.

We develop a deep learning computer vision model to score images according to their aesthetic value and we validate its output through crowdsourcing. By applying it to over 15B Flickr photos, we study for the first time how image beauty is distributed over a large-scale social system. Beautiful images are evenly distributed in the network, although only a small core of people get social recognition for them. To study the impact of exposure to quality on user engagement, we set up matching experiments aimed at detecting causality from observational data. Exposure to beauty is double-edged: following people who produce high-quality content increases one’s probability of uploading better photos; however, an excessive imbalance between the quality generated by a user and the user’s neighbors leads to a decline in engagement. Our analysis has practical implications for improving link recommender systems.

1 INTRODUCTION

The user experience in online communities is mainly determined by the social network structure and by the user-generated content that members share through their social connections. The relationship between social network dynamics and user experience [26, 66], as well as the influence of quality of content consumed on user engagement [9, 10, 24] have been extensively researched. However, the relationship between network properties and the production of quality content remains largely unexplored. This interplay is key to reach a full understanding of the user experience in online social systems. Learning how people engage with a platform in relation with the content they produce and consume is crucial to prevent churning of existing users, keep them happy, and attract newcomers.

The growing availability of interaction data from social media, along with the development of increasingly accurate computational methods to evaluate quality of textual and visual content [31, 33, 37, 41], has recently provided effective means to fill this knowledge gap. We tap into this opportunity and we aim to advance this research direction by providing the first large-scale study on the production and consumption of quality in online social networks.

We do so through three main contributions. First, we develop a new deep learning model able to capture the beauty of a picture ($\S$4), as confirmed by a large-scale human crowdsourcing evaluation ($\S$5). Second, by applying the model to 15B public photos from Flickr ($\S$3), we are able to draw the quality profile of the photo collections uploaded by several million users and to partition these users into coherent classes based on the combination of their connectivity, popularity, and contributed quality. This provides the largest-scale description to date of the distribution of quality in an online community. We explore for the first time the relationship between quality production and network structure ($\S$6). Most importantly, we set up matching experiments aimed at inferring causal relationships from longitudinal data which allows us to learn more about the combined effect of social network connectivity and the process of quality production on user behavior.

Key findings from the analysis include the following:

- Unlike popularity, quality is evenly distributed across the network. The resulting mismatch between talent and attention received leaves large portions of the most proficient users with little peer recognition. Users who produce high-quality content but receive little social feedback tend to stay active only for short periods.

- The level of user-generated quality is correlated with individual social connectivity, which causes a majority illusion effect: users are exposed to images whose average beauty is considerably higher than the average beauty of photos in the platform.

- Users tend to be assortatively connected with others who produce pictures with similar beauty levels to their own. We find that this network property is partly credited to influence (following talented people increases one’s content beauty in the near future) and by the instability of social connections with high imbalance of contributed qualities (users tend to become inactive or churn out if the quality of their neighbors’ photos is substantially higher or lower than their own).

The outcomes of our study have practical implications in the domain of recommender systems. We sketch a simple proof-of-concept of a social link recommender algorithm that maximizes the beauty flow while limiting the beauty imbalance between friends ($\S$7). Simulations show that this simple strategy balances beauty supply
and demand, increasing the level of social inclusion in the class of talented yet unpopular users.

2 RELATED WORK

Computational Aesthetics. With this work, we build on recent literature exploring the possibility of measuring the intrinsic visual quality of images. Previous related work belongs to the research field of computational aesthetics, a domain in which computer vision is used to estimate image beauty and quality. Traditional aesthetic prediction methods are based on handcrafted features reflecting the compositional characteristics of an image. Datta et al. [14] and Ke et al. [32] were pioneers in this field, with their early work on training classifiers to distinguish amateur from professional photos. Researchers have produced increasingly more accurate aesthetic models by using more sophisticated visual features and attributes [18, 47], looking at the contribution of semantic features [42, 46], and applying topic-specific models [39, 48] and aesthetic-specific learning frameworks [69]. Similar hand-crafted features have successfully been employed to predict higher-level visual properties, such as image affective value [40], image memorability [29], video creativity [53], and video interestingness [30, 52]. Such hand-engineered features are of crucial importance for computer vision frameworks requiring interpretability. Recently, Convolutional Neural Networks (CNNs) have become a very popular alternative to hand-crafted features in the computer vision domain, due to their impressive performance on image analysis tasks [57]. The few pieces of work that tested CNNs for aesthetic scoring have done so on professional image corpora [33, 37, 41]. In this work, we develop a CNN-based aesthetic predictor and compare its performance to existing work and to human evaluation through a crowdsourcing experiment.

Media Content Quality and User Experience. Similar to our work, several user studies in controlled lab settings have evaluated how quality affects user experience in relation to different types of media content. Gulliver et al. [24] found that video frame rate and network characteristics such as bandwidth and video topic impact user perception of information quality. Bouch et al. explored the importance of contextual and objective factors for media quality of service [9], and Ceaparu et al. found causes of user frustration in web browsing, e-mail, and word processing [10]. In this work we explore the impact of visual aesthetic quality in online social networks. Past research has demonstrated the importance of visual aesthetics in improving user satisfaction and usability of web pages [16, 35]. In the context of online advertising, researchers have found that image quality properties can impact the user experience of the ad viewed [72]. Aesthetically appealing preview thumbnails increase the clickthrough probability of a video [63]. In recent work, Schifanella et al. showed how existing features for aesthetics, embedded in topic-specific aesthetic models, can be used to surface beautiful but hard-to-find pictures and that content quality is only weakly correlated with its popularity [60]. We build on such work to analyze how quality production and consumption are related to the social network topology at scale.

Networks and Media Diffusion. Bakshy et al. examined the role of social networks in information diffusion with a large-scale field experiment where the exposure to friends’ information was randomized among the target population [5]. They found that users who are exposed to friends’ social updates are significantly more likely to spread information and do it sooner than those who are not exposed. They further examine the relative role of strong and weak ties in information propagation, showing that weak ties are more likely to be responsible for the propagation of novel information. Social exposure, assortative mixing, and temporal clustering are not the only factors that drive information diffusion and influence. Aral et al. studied the effect of homophily in explaining such evidence [4]. They developed a dynamic matched sample estimation framework to distinguish influence and homophily effects in dynamic networks, and they applied it to a global instant messaging network of 27.4 million users. Stuart addressed the problem of estimating causal effects [64] using observational data, and explained how to design matching methods that replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distribution. Those type of techniques are increasingly used being used to analyze digital traces [2]; we leverage them in our work too.

3 DATASET

Flickr is a popular photo-sharing platform on which users can upload a large number of pictures (up to 1 TB), organize them via albums or free-form textual tags, and share them with friends. Users can establish directed social links by following other users to get updates on their activity. Since its release in February 2004, the platform has gathered almost 90 million registered members who upload more than 3.5 million new images daily1.

We collected a sample of the follower network composed of the nearly 40M public Flickr profiles that are opted-in for research studies and by all the 570M+ following links incident to them. For each profile in the sample, we get the complete information about the photos they upload (around 15B in total), the favorites their photos receive from other users, and the groups they are subscribed to. Every piece of information is annotated with timestamps that enable the reconstruction of the full temporal profile of a user’s public activities. The whole data spans approximately 12 years, starting from the debut of the service in 2004 until March 2016.

The distributions of the main activity and popularity indicators, along with their average values ($\mu$), are shown in Figure 1. As expected, all distributions are broad, with values spanning several orders of magnitude.

4 SCORING IMAGE BEAUTY

The first step towards a complete characterization of aesthetic quality in the Flickr network is to quantify beauty at the image level. To do so, we trained a deep neural network to produce a pixel-based aesthetics score. To boost performance, this network was pre-trained on a large-scale supervised image recognition task, and then the final layers were fine-tuned on our aesthetics estimation task [23].

Training vs Fine-tuning. Deep neural network architectures are essentially layers of artificial neurons that progressively abstract
We start with a network pretrained for object detection. The architecture and training process for this network are similar to the reference model proposed by Krizhevsky et al. [34]. However, we introduce a few fundamental changes. We doubled the size of the fc6 (second-last) layer from 4096 to 8192. We also used a final fc8-layer consisting of 21841 units (instead of 1000), corresponding to the complete collection of annotated objects in the ILSVRC ImageNet dataset [57]. We found that for the purpose of pre-training, predicting all objects was more effective than just using the standard 1000 categories typical in the ILSVRC challenges. This also allowed us to use the complete ImageNet dataset of about 14 million images.

Fine-Tuning on Aesthetic Scoring. After pre-training on the ImageNet classification task, we fine-tune the network for the aesthetic scoring task. The training set for the aesthetic quality classification task is an internal dataset created using a proprietary social metric of image quality based on Flickr’s user interaction data, that has proved to correlate closely with subjective assessments of aesthetic quality. We rank all images from the YFCC100MM dataset [67] according to this metric and then create buckets of “low quality”, “median quality”, and “high quality” by sampling images from the bottom 10-percentile, the middle 10-percentile, and the top 5-percentile respectively. The aesthetic classification task requires the network to assign images to the right quality buckets. We then proceed to fine-tuning, replacing the final layer of the object detection network with the 3-way aesthetic quality classification task. This means that the output layer is made of 3 neurons, one for the low category, one for the medium category, and one the for high quality category. Initially, we fine-tune just the final fully connected layer; after convergence, we fine-tune the whole network.

Network Evaluation. The output layer of the network yields three scores via softmax—these correspond to the probabilities of a photo’s “low” \( p_{LQ} \), “medium” \( p_{MQ} \), and “high” \( p_{HQ} \) quality. Each probability is the output of the corresponding neuron. Collectively, the scores correspond to the output of a softmax function evaluating the categorical probability distribution over the 3 possible outcomes: low, medium, and high. The three scores (in the range \([0,1]\)) sum up to 1. In empirical evaluations, we noticed that the per-class network accuracy is higher for images in the low and high quality categories. We therefore design our continuous scoring formula by considering the output of the neurons corresponding to the low and high classes only, namely \( p_{LQ} \) and \( p_{HQ} \), respectively. We combine these two into a single aesthetic score by subtracting the low quality probability from the high quality probability, followed by normalization to the range \([0,1]\):

\[
s = \frac{1}{2}(p_{HQ} - p_{LQ} + 1) \tag{1}
\]

The network achieves a final single-crop test accuracy of 62.5%, almost twice the accuracy of a random classifier. To further verify the performance of our approach, we compare it with state-of-the-art methods for automatic aesthetic assessment. We fine-tune the network with AVA, one of the most widely-used benchmarking datasets [46]. Following existing work, we re-train the network for binary aesthetic classification, a simpler task compared to the 3-way decision we use, and achieve a classification accuracy of 77.6%, thus in line with the most recent state-of-the-art on the same dataset, which stands between 75% and 79%, depending on the training and test setup [31, 33, 37, 41].

Classification vs. Regression. We tested the possibility to predict a continuous aesthetic score using regression: we obtained a continuous aesthetic score for each sample in our training set by placing the categorical annotations on a continuous scale and normalizing in the range \([0,1]\). We designed the output layer to contain one single neuron predicting the aesthetic score; we trained to minimize Euclidean loss. Although this approach has been found to be effective by Kong et al. [33], we found in empirical evaluations that this approach is less effective than our proposed methodology. As a matter of fact, our accuracy on the AVA dataset (77.6%) is 5 points higher than the regression-based framework proposed by Kong et al. [33] (72% for the regression based on visual data only).
In addition to the standard performance test on benchmarking datasets, we further evaluate the effectiveness of the aesthetic network with a crowdsourcing experiment. We ask people to evaluate pictures in terms of their beauty, and then compare the human judgments to the aesthetic score predicted by our framework. To design our experiment, we draw inspiration from the image beauty assessment crowdsourcing experiments conducted by Schifanella et al. [60].

Crowdsourcing tasks are complex and can be influenced by unpredictable human factors [43]. Modern crowdsourcing platforms offer control mechanisms to tune the annotation process and enable the best conditions to get high-quality judgments. To annotate the beauty of our images, we use CrowdFlower™, a popular crowdsourcing platform that distributes small tasks to online contributors in an assembly line fashion.

**Data selection.** To help the contributor to assess the image beauty more reliably, we build a photo collection that represents the full popularity spectrum, thus ensuring a diverse range of aesthetic values. To do so, we identify three popularity buckets obtained by logarithmic binning over the range of number of favorites $f$ received. We refer to them as tail ($f \leq 5$), torso ($5 < f \leq 45$), and head ($f > 45$) to identify the characteristic segments of the broad distribution. From the validation set used to evaluate the aesthetic network, we randomly sample 1000 images from each bucket. Images from such diverse popularity levels are also likely to take a wide range of aesthetic values, thus ensuring aesthetic diversity in our corpus, typically very important for the crowdsourcing of reliable beauty judgments [51].

**Crowdsourcing task setup.** The task consists in looking at a number of images and evaluating their aesthetic quality. At the top of the page we report a short description of the task and we ask to answer the question “*How beautiful is this image?*” (Figure 2). The contributor is invited to judge the intrinsic beauty of the image and not the appeal of its subject; for example, artistic pictures that capture non-conventionally beautiful subjects (e.g., a spider), should be considered beautiful. Out of all the possible rating scales commonly used in crowdsourcing [22], it has been shown that the 5-point *Absolute Category Rating* (ACR) scale is good way to collect aesthetic preferences [62]. We therefore ask contributors to express their judgments by selecting one out of 5 aesthetic categories from “Unacceptable” to “Exceptional”. To guide the contributor in its choice, two example images for each grade are shown (Figure 3). Examples are Flickr images that have been unanimously judged by three independent annotators to be clear representative instances of beauty grade. Below the examples, the page contains 5 randomly selected images to be rated. The images in each page are randomly selected and displayed in an approximate equally-large size to minimize any skew in the perception of image quality [12, 22].

**Quality control.** To maximize the quality of human judgments, we apply several controls on the contributors’ input. First, we open the task only to Crowdflower contributors with an “excellent” track record on the platform (responsible for the 7% of monthly CrowdFlower judgments). We also limit the task to contributors from specific countries (see Table 1), to ensure higher cultural homogeneity in the assessment of image beauty [17, 25, 45, 58, 70]. Second, we cap the contributions of each worker to a maximum of 500 judgments to prevent potential biases introduced by the predominance of a small group of active workers. Last, we discard all the judgments of contributors who did not annotate correctly at least 6 out of 8 Test Images that are presented to them in an initial Quiz page and randomly throughout the task, disguised as normal units. Similar to the examples, Test Images are Flickr pictures that have been unanimously judged by three annotators to be clear representative instances of a beauty score.

**Agreement.** Each photo receives at least 5 judgments by as many independent contributors. Despite aesthetics assessments having a strong subjective component, we register a good level of agreement between annotators, in line with previous work on image beauty [60]. The average percentage of matching annotations over 5 judgments is 73%. When judgments do not match exactly, they usually cluster around two consecutive scores; the average standard deviation around the average score is 0.45, less than half point. In alternative to matching, we also compute Cronbach’s $\alpha$, a widely-adopted metric to assess inter-rater agreement on aesthetics tasks [62]. The Cronbach’s coefficient is 0.77, a value that falls in a range that is commonly considered a Good level of inter-rater consistency [7].

**Results.** Having collected reliable annotations on 3,000 validation images, we test the aesthetic network predictions relative to the ground truth as follows. We are interested in a predicted score that, regardless of its range or distribution, preserves the ranking of the original beauty scores assigned by human annotators. To check that, we compute the Spearman rank correlation coefficient $\rho$ between the predicted score and the crowdsourced score. We find

![Figure 2: Screenshot of the crowdflower job: instruction examples (left) and voting task (right).](http://www.crowdflower.com/)
a high correlation $\rho = 0.48$ (with $p < 0.01$), which suggests that our automatic aesthetic scoring method is an effective proxy of human aesthetic judgment. To further dig into this intuition, we partition the validated images into 10 equally-spaced intervals of predicted aesthetic score (i.e., $[0, 0.1], ..., [0.9, 1]$). We then compute the average crowd-sourced beauty score for all images in each bucket. Figure 4 shows that the average crowdsourced score linearly increases with the predicted beauty decile, further confirming that our aesthetic framework performs comparably to human evaluation on this task.

Additionally, we test the level of agreement of the algorithmic beauty prediction with the judgment of human labeler using the state of the art approach proposed by Ye et al. [71]. Their evaluation method, inspired by the work on consensus methods by Dawid and Skene [15], has been used to assess the robustness of crowdsourced affective data and can be used to estimate how much machine-generated labels can accurately mimic the human judgments. We apply the method to the human-generated beauty scores and found an average reliability score of $\tau = 0.71$, with peaks of $\tau_{\text{max}} = 0.95$, much higher than the reliability of a random annotator $\tau_{\text{rand}} = 0.22$ (to obtain this number, we added to the pool of annotators a fake annotator giving random judgments). Next, we re-scale the continuous scores predicted by the aesthetic network over a discrete 5-point scale, in order to make machine predictions comparable to human labels. We add the scaled predictions to the previous list of judgments by treating the machine-generated scores as the output of an additional annotator. We re-calculate reliability of all annotators, including the machine: we find that the reliability of the machine judgments stands at 0.77, in line with the average reliability score.

### 6 NETWORK EFFECTS

While previous work has studied beauty at the picture level, our large-scale rating of image quality further enables us to analyze how beauty is produced over a large social network. In the following, we will characterize the beauty $b(i)$ of a user $i$ as the average beauty of all of $i$’s public photos. We will refer to this score as user **beauty** or user **quality**, for brevity. When time is relevant to the analysis, we will use $b_{\text{t}}(i)$ to denote the average beauty of pictures posted by user $i$ during week $t$ and $b^t(i)$ to denote $i$’s photos average beauty until week $t$. Although summarizing the quality production of a user with a single indicator is limiting, it helps to simplify the analysis that follows. In future work we plan to consider more complex quality profiles that include, for example, the variance of photo quality.

Unlike the heavy-tailed distributions of activity and popularity indicators (Figure 1), the user beauty is bell-shaped distributed, with a slightly heavier right tail (Figure 5, left). This leads to a mismatch between the ability to produce high-quality content and the social attention received by the community. As a result, we observe a more marked inequality in the distribution of the average number of favorites per photos across users than in the distribution of average user quality (as measured by the Gini index, Figure 5, right). This finding is in line with previous work on a smaller data sample [60] showing that high-quality Flickr pictures are distributed across different ranges of popularity.

In other words, this heavy imbalance reveals that a large number of users who post high-quality photos receive very little social attention. Next, we map the average user beauty on the Flickr
follower network to further investigate the unexplored relationship between user beauty and social connectivity patterns. In particular, we are interested to shed light on two unexplored matters: i) how the quality is distributed over the network (§6.1) and ii) the causal impact that the quality users are exposed to has on their own activity and engagement (§6.2). These issues are crucial to managers of online communities, who aim to provide all users with high-quality content and retain them as long as possible. However, those could not be addressed in the past due to the scarcity of large-scale data suitable for such analysis and the lack of reliable and efficient tools to measure content quality.

6.1 Distribution of quality over the social network

Different activity indicators of social media users tend to be correlated. This has been verified in multiple social media platforms, including nodal degree [44, 59]. We are interested in verifying whether the level of user quality is correlated to social connectivity or other activity indicators.

Q1: Is quality correlated with social connectivity? We compute the Spearman rank correlation $\rho$ between user beauty and nodal degree. We find A1: a small but positive correlation $\rho$ between user beauty, indegree ($\rho = .22$), and outdegree ($\rho = .24$) (Figure 6). Beauty is also weakly associated with the average number of favorites received by a user ($\rho = .17$) and exhibits a slightly negative correlation with the number of photos posted ($\rho = -.03$), confirming that neither content popularity nor volume of contributions are strong determinants of quality.

The association between quality and connectivity can have higher-order effects. In online social networks, as in offline social environments, people lack global knowledge of the overall population’s characteristics, since their view of the external world is mediated by their direct social connections. This local constraint might lead to an over-representation of some rare population attributes in local contexts. This phenomenon has been observed in the form of the so-called friendship paradox [20, 27], a statistical property of networks with broad degree distributions for which on average people have fewer friends than their own friends. The paradox has been recently extended by the concept of majority illusion [36], which states that in a social network with broad degree distribution and binary node attributes there is a systematic biased local perception that the majority of people (50% or more) possess that attribute even if it is globally rare. As an illustrative example, in a network where people drinking alcohol are a small minority, the local perception of most nodes can be that the majority of people are drinkers just because drinkers happen to be connected with many more neighbors than the average.

In our context, we are interested in measuring the presence of any skew in the local perception of the quality of user-generated content. So we ask:

Q2: Does the correlation between connectivity and quality creates a majority-illusion effect on user beauty?

To estimate the presence of any local perception skew, we calculate the proportion of users in a node’s neighborhood whose quality

$$b_{nn}(k) = \frac{1}{|\{i : b(i) = k\}|} \sum_{i : b(i) = k} \frac{\sum_{j \in \text{out}(i)} b(j)}{|\text{out}(i)|}$$

(2)

Figure 6: Average user beauty for users with fixed indegree (left) and outdegree (right). The positive slopes (Spearman correlations $\rho = .22$ and $\rho = .24$, respectively) indicate that users who are more connected tend to produce higher-quality content. 95% confidence intervals are shown.

is above the global average quality of users in the network ($\mu = 0.26$, as per Figure 5 (left)), and compare it with the actual portion of users in the overall population with beauty above the global average\(^*$). We find that A2: the majority illusion holds in our data sample. Overall, 43% of the users typically produce content with above-average quality; however, 65% of the population has more than 43% of their friends with above-average quality. The phenomenon is very strong for the nearly 20% of users who have more than 86% of their neighbors falling into this category (double or more than what is expected). Nevertheless, the majority illusion does not imply that people preferentially connect to very talented users. Next, we investigate the relationship between the beauty levels of connected individuals.

Q3: Are social connections established between users with similar beauty?

A typical pattern found in several ecological and social networks is assortative mixing, namely the high likelihood of nodes to be connected to other nodes with similar properties. This propensity is gauged with the correlation spectrum [6], a measure that puts in relation all the nodes that have a fixed value $k$ of a target indicator with the average value of the same indicator of their neighbors. By setting user beauty as the target indicator, we measure the correlation spectrum by computing the average neighbor beauty of all those users with a fixed user beauty $b = k$, for all possible values of user beauty

$\bar{b}_i$ is user i’s beauty and $\Gamma_{\text{out}}(i)$ is i’s out-neighbors. Figure 7 shows the trend of $b_{nn}$ for all possible values of $k \in [0, 1]$, obtained by partitioning the beauty range into 100 equally-sized bins. The positive slope of the curve (Spearman correlation $\rho = .48$) reveals an assortative trend, which indicates that A3: users tend to be linked to accounts that publish photos with similar quality as their own. The trend is particularly clear for users with beauty in the range $[0.08, 0.6]$, which corresponds to 90% of our sample’s population. To tell apart any assortativity trend from statistical artifacts, we need to compare the results obtained on the real data.

\(*\)Because the overall distribution of quality has a shape that is close to normal, the results do not change considerably when using the median instead.
Q4: Is the user beauty affected by the content produced by their social neighbors? The quality of content produced by users might be affected by the quality of the content that their social contacts produce. In particular, we hypothesize that, on average, the user beauty increases as an effect of the creation of a new social connection with a higher-beauty user.

Q5: Does a heavy quality imbalance between connected individuals affect their social engagement? We hypothesize that, on average, heavy imbalance between the user beauty and the average beauty of its neighbors leads to a drop in engagement. This intuition is backed by one of the core principles of the Social Exchange Theory [8], which states that reciprocity is necessary to maintain a stable social relationship. When reciprocity fails consistently, at least one of the parts is likely to withdraw. In online social platforms, users join with specific expectations; when those are not met, the likelihood of abandonment is expected to rise. Specifically in the context of Flickr, talented photographers won’t feel their efforts being reciprocated if the quality of all other contributors’ content is mediocre, whereas casual photographers might feel overwhelmed if mostly surrounded by professionals and will more likely regress to a lurking state or even unsubscribe.

6.2 Network effects on user retention and quality production

6.2.1 Quality, network, and engagement. The assortative mixing of quality in the social network could be ascribed mainly to homophily or influence [1]. On one hand, users might preferentially connect to accounts that publish pictures with a similar quality to their own. This would seem natural in a platform like Flickr that hosts a heterogeneous user-base: semi-professional photographers might be interested in following users who are well-versed in the use of photographic techniques, whereas casual users might be following each other mostly for social reasons, unconcerned about aesthetic photo quality. On the other hand, pairs of users might be imbalanced in terms of their quality at the time they connect and close their quality gap later on, over time. For example, amateur photographers could follow professionals and learn new skills from them, thus improving the quality of their pictures.

The interplay between homophily, influence, and other factors leading to assortative mixing has been the subject of a number of studies [3, 13] that explored these phenomena on a wide range of user attributes (e.g., demographics, topical preferences). However, despite its crucial role in growing and maintaining user engagement [19], content quality has never been investigated in relation to such network properties. We aim to shed light on this relation.

Figure 7: Correlation spectrum of user beauty on the Flickr follower network. The highlighted interval on the beauty axis includes the user beauty values of 90% of the population. Variance is shown. The correlation spectrum for a null-model with randomly reshuffles user beauty scores is reported for the sake of comparison.

with a suitable null model. When using a null model that randomly reshuffles the beauty values between all users, keeping unchanged their social connections, the trend is lost.

In summary, we have found that user quality correlates with individual connectivity, which in turn leads to a majority illusion phenomenon, where high-quality users are more visible than lower quality ones. Also, beauty is an assortative property, with user being preferentially connected to others with similar quality.

6.2.2 Matching experiments for causal inference. To answer the two questions above, we set up matching experiments aimed at inferring causality from the observational data. In natural experiments, estimating the statistical effect of a treatment on a population can be done through randomization. Provided that the population is sufficiently large, randomly allocating individuals across the treatment and control groups cancels the potential biases by equalising all the observable factors as well as unobserved variables that have not been explicitly accounted for. Without the possibility to run controlled experiments over the Flickr user-base, we need to infer causality from observational data. That is a much harder task [4, 61] because the benefit of randomization is lost, as the set of individuals who received the treatment is often pre-determined.

Matching experiments provide a way to reliably estimate the statistical effect of a treatment on a dependent variable from longitudinal data. The key intuition is to match the treated group \(G_t\) with a control group \(G_c\) whose members did not receive the treatment and are statistically indistinguishable (i.e., only marginally different) from the treated group on all observable covariates.

There are several ways to perform matching [50, 55, 64] and to measure the equivalence between treatment and control groups. Here we borrow a framework introduced by Rubin [56] and later summarized by Stuart [64], which has been successfully used in other observational studies aimed at inferring causality [2]. This framework assumes that \(G_t\) and \(G_c\) are somehow formed and provides a function to check their statistical equivalence. The two groups are said to be balanced on a covariate \(X\) when the covariate’s standardized bias \(SB\), namely the difference of its mean values \(\bar{X}\) in the two groups divided by the standard deviation \((\sigma)\) in the treated group, is under a given threshold commonly set to 0.25. Formally:

\[
SB_X(G_t, G_c) = \frac{\bar{X}_t - \bar{X}_c}{\sigma(X)} \leq 0.25
\]  

(3)

The groups are overall balanced—and therefore indistinguishable, from a statistical point of view— only if they are balanced on all their covariates.

Algorithm to balance treatment and control groups. Given a treatment group \(G_t\), we set a greedy iterative procedure to select a
corresponding balanced control group $G_c$. At step 1, a candidate control group $G_c^1$ such that $|G_c| \gg |G_t|$ is selected from the set of non-treated units. At step $n$, the standardized bias $SB(G_t, G_c^n)$ is computed for every covariate. For all the covariates that do not satisfy the balance constraint, we remove from the control group the elements that most contribute to the mismatch. Specifically, we cut off the 1% of experimental units with the highest values of the covariate, when $SB$ is negative, or with the lowest values, when $SB$ is positive. At each iteration, further pruning could be required on different sets of covariates. The algorithm stops when the condition $SB(G_t, G_c^n) \leq 0.25$ is satisfied by all the variables. The procedure does not have a theoretical guarantee to stop before pruning out all the elements of $G_c$, in which case the algorithm should be restarted with a different seed control group. In our experiments we always observe convergence before $|G_c| < |G_t|$.

Next, we describe how this framework is instantiated on our Flickr data. For these experiments, we have considered only users with at least 10 outgoing social links (i.e., followees) and who have uploaded photos in at least 12 distinct weeks (which implies they all have at least 12 photos each). This filtering step yielded a subset of 2.7M users.

### 6.2.3 The effects of neighbors’ beauty

**Beauty inspires beauty.** To answer question Q4, we use link creations as events to split users between treatment and control groups. We include in the treatment group users who have created a social link towards accounts with higher quality than their own and compare them with a control group whose members have connected to users with equal or lower quality.

Operationally, we partition the timeline of events in our data into discrete slots of one week each. For each week $w$, we iterate over the set of users $U_w$ who have been active during that week and have been active for at least 12 non-consecutive weeks before it. All users who added at least one link towards higher-quality users on that week are added to the control group $G_t$. Among the remaining users in $U_w$, we add to $G_c$ those who created any number of links during that week. Each element in the two groups is described with a vector of covariates that accounts for all the main aspects related to the popularity, activity, age, and quality of the users and to the quality and activity of their neighbors, measured at the beginning of week $w$ (Table 1). As we iterate over all the weeks in the timeline, users performing link creations during several weeks will be added multiple times to any of the two groups. This is acceptable from an experimental design perspective [64]: two versions of the same user profile at different times will have different vector of covariates, thus we will effectively consider them as two distinct user instances.

After the two groups are built, we execute the algorithm described in the previous section (§6.2.2) to obtain two statistically balanced groups. The matching algorithm yielded a pair of balanced groups with $SB < 0.25$ for all covariates and an average $SB$ of 0.18. We then compare the two groups on an outcome variable that reflects our research question. For every user instance $i$ in $G_t$ or $G_c$, we measure the quality variation of its produced content after the link creation event. This is done by computing the ratio $\Delta_b k$ between the beauty of the user’s photos uploaded in the week after the link creation ($b^{w+1}(i)$) and the average beauty of all its photos posted prior to the link creation event ($\bar{b}^w(i)$). When averaged over all the elements of the group, the outcome variable is defined as follows:

$$\Delta_b k(G_t) = \frac{1}{|G_t|} \sum_{i \in G_t} \frac{b^{w+1}(i)}{\bar{b}^w(i)}; \text{(same for } G_c).$$

Figure 8 depicts a simplified sketch of the matching experiment. The measure of $\Delta_b k$ confirms our hypothesis: the treatment group experiences an average 2% increase in $\Delta_b k$, whereas no significant increase is found in the control group (Figure 9 left).

Using the same matching setup, we run two additional experiments with new pairs of groups. First, to assess how much the influence effect is augmented by the number of new connections, we run another matching experiment that includes in $G_t$ only users
We find that, the greater the beauty differential, the greater the confidence intervals are shown.

Last, to measure how much the average neighbor beauty deviates from the user beauty

\[ \hat{b}^w(\Gamma^w(i)) = \frac{1}{|\Gamma^w(i)|} \sum_{j \in \Gamma^w(i)} \hat{b}^w(j), \]

where \( \delta \) is the difference in treatment over control. We observe that the treatment group has higher probability of inactivity that grows from +5% to +20% in the first 12 weeks (Figure 11). In conclusion, we find that A5: people exposed to photos that deviate too much, in terms of quality, from their own contributed content, are more likely to become disengaged in the future.

### 7 Beauty-Based Link Recommender

Classic link recommendation approaches based on the graph structure (e.g., common neighbors and all its variations) tend to suggest popular and very connected users [65], thus increasing the link-age to—and consequently the level of attention on—already well-regarded individuals, keeping potential new talents away from the spotlight. However, since connectivity and user quality are largely orthogonal, algorithms that favor highly-connected users won’t necessarily provide adequate visibility to high-quality content.

This point is made particularly evident if we group users according to the combination of their popularity and produced quality. We cluster Flickr users according to three variables: content quality (average beauty of the user’s photos), popularity (average number of favorites per photo), and connectivity (number of followers). Given the diversity in terms of range and distributions of such variables, we first log-transform their values and then normalize them to the range [0, 1]. Next, to identify groups of photographers with similar characteristics, we use K-means clustering over these dimensions. We vary K from 2 to 10, and select K = 4 according to...
Table 2: Clustering results. Photographers are divided into four groups based on their quality, popularity, and connectivity. The normalized values of those three dimensions for the four centroids are reported.

<table>
<thead>
<tr>
<th>Centroid</th>
<th>%users</th>
<th>beauty</th>
<th>fav/photo</th>
<th>connects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low quality</td>
<td>41.2%</td>
<td>0.17</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Forlorn beauty</td>
<td>28.1%</td>
<td>0.42</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Regular user</td>
<td>22.1%</td>
<td>0.25</td>
<td>0.01</td>
<td>0.21</td>
</tr>
<tr>
<td>Superstar</td>
<td>8.6%</td>
<td>0.42</td>
<td>0.15</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 3: Average value of descriptive metrics for users in different clusters

<table>
<thead>
<tr>
<th>Metric</th>
<th>Low</th>
<th>Forlorn</th>
<th>Regular</th>
<th>Superstar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo count</td>
<td>1060</td>
<td>200.4</td>
<td>1869</td>
<td>822.4</td>
</tr>
<tr>
<td>Time on platform</td>
<td>104.4</td>
<td>84.68</td>
<td>187.0</td>
<td>198.3</td>
</tr>
</tbody>
</table>

the gap statistic [68]. The cluster centroids are reported in Table 2. Four classes of users emerge:

(1) **Low Quality**: The biggest cluster contains almost half of the users. It corresponds to the long tail of “beginner photographers” who produce average-to-low quality content, with limited activity and low connectivity in the network.

(2) **Forlorn Beauty**: The second biggest cluster gathers excellent photographers (highest average beauty value among the clusters considered) who receive very little attention from other Flickr users.

(3) **Regular Users**: The regular semi-professional photographer on Flickr, sharing average-to-high quality pictures. These users are characterized by a moderate popularity within the network.

(4) **Flickr Superstars**: The smallest cluster groups together all those professional photographers (beauty level similar to the Forlorn Beauty cluster) who are the foundation of the Flickr network, with many favorites and followers. Typically, these Superstars are the ones who appear in showcase pages such as the Flickr Explore.

The clustering results confirm that the talent of a large portion of the user-base—more that 1/4th of the overall population—remains largely untapped, despite its high skill level (as evidenced by the high average beauty value). This group of users is associated with a lower time on platform, measured as the number of weeks with at least one photo upload (Table 3). This gives further support to the intuition that photographers who do not receive adequate recognition for their contributed value tend to churn out sooner. Furthermore, their activity in terms of number of pictures uploaded is limited (the lowest compared to other user classes), thus reducing the flow of incoming high-quality content in the platform.

Link recommender systems oblivious to quality will disproportionately recommend Superstar users because they are very popular and well-connected. By doing so, users will be exposed to new appealing pictures because recommended contacts produce beautiful photos on average. However, this strategy has two major limitations. First, it reinforces the rich-get-richer phenomenon, depriving the users in the Forlorn beauty class of the attention they deserve by directing it all to the small core of popular users. Last, it worsens the risk of very imbalanced connections: users who post lower-quality pictures will be mainly recommended contacts with considerably higher beauty. This is an undesirable outcome because, as we have shown earlier, accumulating many unbalanced connections increases the risk of inactivity and churn-out.

Next, building on our previous findings, we contribute to address these limitations by sketching a simple link recommendation strategy that (i) rebalances the distribution of attention to give recognition to valuable contributors otherwise forgotten, and (ii) increases the chances of a user to access new high-quality content without aggravating the quality imbalance between producers and consumers, which might cause engagement to drop in the long term.

To test this idea, we simulate a link recommendation task. We compare a classic friend-of-friend approach that recommends the contact with the highest number of common neighbors (CN) with an alternative, quality-oriented algorithm that recommends the user at network distance 2 with the highest average beauty score (BB\textsubscript{2,10}) that is within a small range from the user beauty of the recommendation recipient (±10%), in order to avoid quality imbalance. We simulate both approaches on a random sample of 400K photographers; each of them receives only one recommendation from each approach.

Let us define \( u \) as the generic user who receives the recommendation, \( r \) the recommended contact, and \( R \) the list of recommendations (\( u, r \)). We compare the two approaches on the four indicators listed below.

- **Average user beauty of recommended contacts** 
  \[ b_{recs} = \frac{1}{|R|} \sum_{(u,r) \in R} b(r). \]

- **Average ratio between the user beauty of the recommendation recipient and the user beauty of the recommended contact** 
  \[ b_{ratio} = \frac{1}{|R|} \sum_{(u,r) \in R} \frac{b(u)}{b(r)}, \]
  a value of \( b_{ratio} \) closer to 1 means a lower beauty imbalance since the two quantities in the fraction are closer.

- **Average number of favorites of recommended contacts** 
  \[ f_{avg_{recs}} = \frac{1}{|R|} \sum_{(u,r) \in R} f_{av}(r). \]

- **Proportion of users in the Forlorn Beauty cluster in the recommendation list** 
  \[ p_{forlorn} = \frac{\sum_{u \in R} (r \in ForlornBeautySet)}{|R|}, \]
  where \( I(\bullet) = 1 \) if the condition of its argument is true, 0 otherwise.

Following the findings from §6, we want to keep such imbalance low to avoid user churns on the long term. Moreover, having a higher ratio of Forlorn Beauty users in \( p_{forlorn} \) increases the exposure and potentially the future engagement of these high-quality photographers with little social attention. Figure 12 shows the results of the two approaches.

The CN approach selects contacts with beauty higher than BB\textsubscript{2,10}, but only slightly higher, considering the strict ±10% constraint in BB\textsubscript{2,10}, which, by definition, will limit the maximum level of beauty for new contacts of a given node. On the other hand, CN’s recommended users are 5 times more popular in terms of number

---

1https://www.flickr.com/explore
of favorites. This introduces higher beauty imbalance (+2.35%) and does not reach 
Forlorn users as effectively: $BB_{\pm10}$ suggests 49% more users in that class, comparatively. 
Even though such a simple algorithm is far from being a production-ready solution, the simulation experiment provides initial evidence that better balance in the content consumption dynamics could be easily introduced by complementing current systems with quality-based rules.

8 DISCUSSION AND CONCLUSIONS

Using a novel deep learning computer vision model trained on a vast image corpus from Flickr, we have conducted the first large-scale study on the relationship content quality with the social network structure.

8.1 Implications

Adopting popularity-driven policies to promote content and users in social networks is a fallacious way of growing healthy online communities [11]. Nevertheless, for several years popularity has been one of the core elements of several online services including search, promoted content, and recommendations. For the first time, we have shown that it is possible to run at scale a reliable profiling of users that captures their contributed quality rather than their popularity. This can have direct practical impact not only in recommender systems, but in any application that need to retrieve, rank, or present images. Furthermore, our study about the notion of quality in combination to the network structure yields important theoretical implications in the domain of social network analysis and, more broadly, network science. We have shown that social relationships are not only homophilous, but tend also to be balanced in terms of the quality that the two endpoints produce. In line with the principles of the Social Exchange Theory, we provide empirical evidence that users who entertain strongly imbalanced social relationships in terms of the quality produced increase the risk of becoming inactive or churn out in the future. As we have shown in a simple proof of concept, next-generation link recommender systems could easily factor in the notion of quality imbalance to foster the creation of longer-lasting social ties.

8.2 Biases

The outcome of both the annotation task and the automated beauty scoring can be influenced by several types of biases.

We have developed the aesthetic scoring system by fine-tuning an existing neural network used for object detection. This choice is justified by computational efficiency, has been adopted in previous work, and complies with photographic theory on subject-specific aesthetic rules. Even though the image set we use to train our neural network is very large and diverse in terms of subjects, quality, and photographers, it may still contain biases that could be smoothened out by extending the training phase to multiple datasets of different nature. In future work we plan to conduct a more systematic evaluation of the biases that this approach might introduce when scoring pictures of different subjects.

The evaluation of image quality through online crowdsourcing might be affected by a number of unconscious biases originating by the personal and cultural background of the raters, the way the interface is presented, and the different subjects depicted in the photos. Although we have used a state-of-the-art framework to account for all these potential problems, a more thorough study focusing on residual biases would be desirable.

8.3 Limitations and future work

Our analysis scratches only the surface of this mostly unexplored research area.

Our causal inference analysis groups together similar users to get a balanced matching between control and treatment sets. That is convenient to measure causal effects globally but does not directly allow for a fine-grained analysis of how meaningful user groups (e.g., newcomers vs. professional users) are impacted. The extent to which the exposure to content quality has a different impact on those user categories is an interesting extension of this work.

The deep learning algorithm we use is very powerful but lacks explainability: in contrast with classic image aesthetic frameworks based on compositional features, it is not possible to determine why a picture has a given beauty score. Research in explainability in deep learning is still at an early stage, also in the sub-field of image aesthetics. Expanding the ability of our method to provide human-readable explanations of the beauty score is part of our planned future work.

We have described user quality with a single numeric indicator; multidimensional descriptors could add nuances to the characterization. We have studied the effect of link creation and nearest neighbors on the process of quality production; exploring a wider range of social structures and events could lead to further findings. Our experiments can determine the cause of some network dynamics (e.g., lower user engagement) but cannot provide reliable explanations about why those changes occur; further investigation, possibly including qualitative methods, could provide more clarity on these dynamics. Last, our experimental setting unveils causality but it is not flexible enough to reveal changes in user quality over long periods of time. Our matching strategy is effective in comparing the effect of an event (e.g., link creation) on outcome variables measured right after the event occurs, but is not designed to study long-term effects. Even though treatment and control groups are checked to be statistically equivalent over all covariates at time $t$, the likelihood that their equivalence is preserved after $t$ drops as time passes and this is why, to draw meaningful causal conclusions, it is safe to study only those outcomes (e.g., variation of user beauty) that occur right after $t$. As a direct consequence, it becomes hard to provide a tangible interpretation in terms of user perception of
some of the small, yet significant, short-term influence effects we have found (e.g., +2% in produced photo quality). In the future, we aim at applying more complex frameworks that can provide reliable causal inference on longer time spans.

Despite such limitations, we hope our work contributes to a better understanding of the evolutionary dynamics of social ecosystems.

REFERENCES


[11] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi. 2010. Understanding of the evolutionary dynamics of social ecosystems. Despite such limitations, we hope our work contributes to a better understanding of the evolutionary dynamics of social ecosystems.


[19] Judith Alice Redi, Tobias Hoßfeld, Pavel Korshunov, Filippo Mazza, Isabel Povoa, Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and RichardD. Walk and Jr. Pick, HerbertL. (Eds.). Perception and Per-

[53] Miriam Redi, Neil O’Hare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In CVPR. IEEE.


[60] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. 2015. An Image Is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures. In ICWSM. AAAI.


