Beautiful and Damned. Combined Effect of Content Quality and Social Ties on User Engagement

Luca Maria Aiello, Rossano Schifanella, Miriam Redi, Stacey Svetlichnaya, Frank Liu, and Simon Osindero

Abstract—User participation in online communities is driven by the intertwinement 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 15 B 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.

Index Terms—Content quality, image aesthetics, network effects, causal inference, influence, matching, flickr

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 [1], [2], as well as the influence of quality of content consumed on user engagement [3], [4], [5] 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 [6], [7], [8], [9], 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 (Section 4), as confirmed by a large-scale human crowdsourcing evaluation (Section 5). Second, by applying the model to 15 B public photos from Flickr (Section 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 (Section 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

- *L.M. Aiello and M. Redi are with Nokia Bell Labs, Cambridge CB3 0FA, United Kingdom. E-mail: {luca.aiello, miriam.redi}@nokia-bell-labs.com.*
- *R. Schifanella is with the University of Turin, Torino 10124, Italy. E-mail: schifane@di.unito.it.*
- *S. Svetlichnaya, F. Liu, and S. Osindero are with Flickr, Sunnyvale, CA 94089. E-mail: {stacey, frank, simon}@yahoo-inc.com.*

Manuscript received 21 May 2017; revised 18 Aug. 2017; accepted 22 Aug. 2017. Date of publication 31 Aug. 2017; date of current version 3 Nov. 2017. (Corresponding author: Luca Maria Aiello.)
Recommended for acceptance by L. B. Holder.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2747552

1041-4347 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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 (Section 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. [10] and Ke et al. [11] 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 [12], [13], looking at the contribution of semantic features [14], [15], and applying topic-specific models [16], [17] and aesthetic-specific learning frameworks [18]. Similar hand-crafted features have successfully been employed to predict higher-level visual properties, such as image affective value [19], image memorability [20], video creativity [21], and video interestingness [22], [23]. 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 [24]. The few pieces of work that tested CNNs for aesthetic scoring have done so on professional image corpora [6], [8], [9]. 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. [5] 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 [3], and Ceaparu et al. found causes of user frustration in web browsing, e-mail, and word processing [4]. 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 [25], [26]. In the context of online advertising, researchers have found that image quality properties can impact the user experience of the ad viewed [27]. Aesthetically appealing preview thumbnails increase the clickthorough probability of a video [28]. 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 [29]. 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 [30]. 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 [31]. 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 [32] 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 [33]; 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 daily.[1]

We collected a sample of the follower network composed of the nearly $40\ \mathrm{M}$ public Flickr profiles that are opted-in for research studies and by all the $570\ \mathrm{M}+$ following links incident to them. For each profile in the sample, we get the complete information about the *photos* they upload (around 15 B 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

1. This figure includes public and private photo uploads—http://bit.ly/1LjaTBT

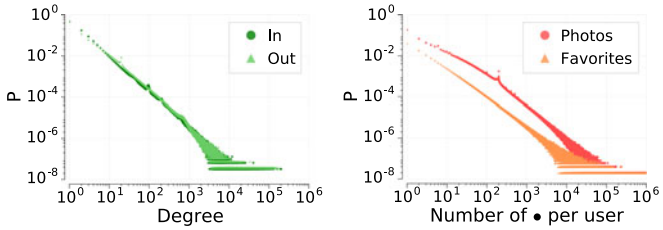

Fig. 1. *Left:* Degree distributions ($\mu_{in} = 19$, $\mu_{out} = 21$). *Right:* Distribution of number of photos uploaded ($\mu = 350$) and number of favorites received ($\mu = 47$). Nearly 80 percent of users receive no favorites.

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 (μ), are shown in Fig. 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 [34]

Training versus Fine-Tuning. Deep neural network architectures are essentially layers of artificial neurons that progressively abstract the input data (the image pixels) into an output network response (the predicted category of the input image). In the training phase, network parameters are tuned in order to maximize metrics such as category prediction accuracy. Given the number of parameters involved in such complex architectures, effectively training neural networks is typically a long, expensive process. A common practice used to speed-up the training process is called fine-tuning, where the last layers of a trained network are modified and re-trained for a new task. In addition to making training more efficient, fine-tuning enables knowledge transfer from the original training data to the new task, improving overall performance. In our case, we start with a network designed for object detection, and then fine-tune it for the task of aesthetic scoring. This allows the aesthetic network to retain some information about the semantic nature of the objects depicted in the image, thus making the system aware of the subject depicted, which is crucial to the correct assessment of a picture's aesthetic value. As a matter of fact photographic theory [35] shows that different aesthetic criteria apply to different subjects: for example, specific photographic techniques should be used when taking pictures with human subjects [36]. Such observations were confirmed by several research works in computational aesthetics [9], [37], [38], [39], which showed that subject-aware aesthetic scorers outperform traditional subject-agnostic aesthetic frameworks.

Training on Object Detection. We start with a network pre-trained for object detection. The architecture and training process for this network are similar to the reference model proposed by Krizhevsky et al. [40]. However, we introduce a few fundamental changes. We doubled the size of the *fc6* (second-last) layer from 4,096 to 8,192. We also used a final *fc8*-layer consisting of 21,841 units (instead of 1,000), corresponding to the complete collection of annotated objects in the ILSVRC ImageNet dataset [24]. We found that for the purpose of pre-training, predicting all objects was more effective than just using the standard 1,000 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 [41] 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 percent, 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 [15]. 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 percent, thus in line with the most recent state-of-the-art on the same dataset, which stands between 75 and 79 percent, depending on the training and test setup [6], [7], [8], [9].

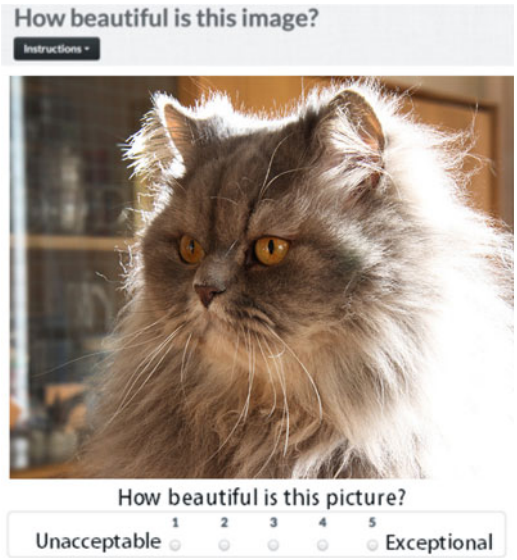

Fig. 2. Screenshot of the crowdflower job: Instruction examples (left) and voting task (right).

Classification versus 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. [8], 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 percent) is 5 points higher than the regression-based framework proposed by Kong et al. [8] (72 percent for the regression based on visual data only).

5 Crowdsourcing Beauty Assessment

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. [29].

Crowdsourcing tasks are complex and can be influenced by unpredictable human factors [42]. 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,[2] 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 1,000 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 [43].

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?"* (Fig. 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 [44], it has been shown that the 5-point *Absolute Category Rating* (ACR) scale is good way to collect aesthetic preferences [45]. 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 (Fig. 3). Examples are Flickr images that have been unanimously judged by three independent annotators to be clear representative instances of that 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 [44], [46].

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 percent of monthly CrowdFlower judgments). We also limit the task to contributors from specific countries,[3] to ensure higher cultural homogeneity in the assessment of image beauty [47], [48], [49], [50], [51]. 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 [29]. The average percentage of matching annotations over 5 judgments is 73 percent.

2. http://www.crowdflower.com/

3. Australia, Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Poland, Spain, Sweden, United Kingdom, and United States.

Fig. 3. Examples of images ranging from beauty score 1 (leftmost) to score 5 (rightmost). These and other examples were provided to crowdworkers for the sake of training.

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 α, a widely-adopted metric to assess inter-rater agreement on aesthetics tasks [45]. 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 [52].

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 ρ between the predicted score and the crowdsourced score. We find 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., $[0, 0.1], \ldots, [0.9, 1]$). We then compute the average crowd-sourced beauty score for all images in each bucket. Fig. 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. [53]. Their evaluation method, inspired by the work on consensus methods by Dawid and Skene [54], 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 $\overline{\tau} = 0.71$, with peaks of $\tau_{max} = 0.95$, much higher than the reliability of a random annotator $\tau_{rand} = 0.22$ (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 beauty further enables us to analyze the how beauty is produced over a large social network. In the following, we will characterize the beauty $\bar{b}(i)$ 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 relevat to the analysis, we will use $b^t(i)$ to denote the average beauty of pictures posted by user i during week t and $\bar{b}^t(i)$ 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 (Fig. 1), the user beauty is bell-shaped distributed, with a slightly heavier right tail (Fig. 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

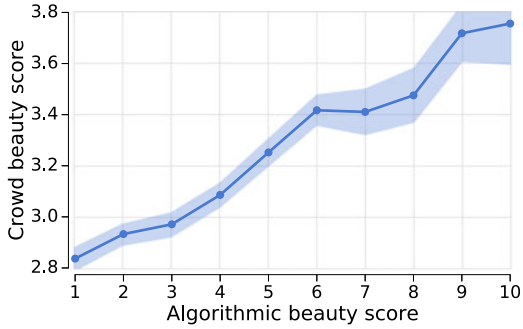

Fig. 4. Average beauty as assessed by crowdworkers against the algorithmic beauty from our deep learning model. Spearman correlation $\rho = 0.48$. Ninety-five percent confidence interval is shown.

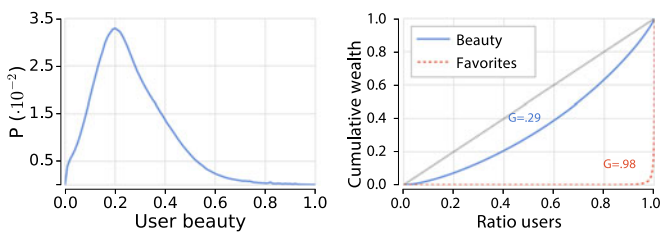

Fig. 5. Distribution of beauty scores; $\mu = 0.26$ (left). Inequality of resource distribution (average beauty and average favorites) across users visualized with the Lorenz curve. Gini coefficients: $G_{favs} = 0.98$, $G_{beauty} = 0.29$ (right).

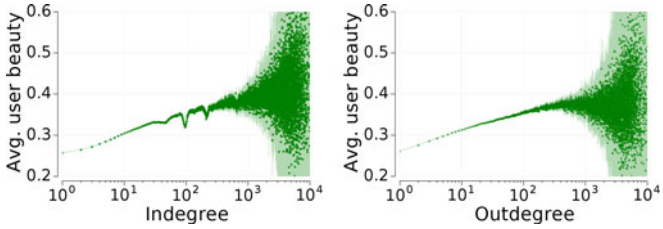

Fig. 6. Average user beauty for users with fixed indegree (left) and outdegree (right). The positive slopes (Spearman correlations $\rho = .22$ and $\rho = .24$, respectively) indicate that users who are more connected tend to produce higher-quality content. Ninety-five percent confidence intervals are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

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, Fig. 5, right). This finding is in line with previous work on a smaller data sample [29] 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 (Section 6.1) and ii) the causal impact that the quality users are exposed to has on their own activity and engagement (Section 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 Flickr, on a wide range of indicators, especially in relation to nodal degree [55], [56]. 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 ρ between user beauty and nodal degree. We find *A1: a small but positive correlation ρ between user beauty, indegree ($\rho = .22$), and outdegree ($\rho = .24$)* (Fig. 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* [57], [58], 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* [59], 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 percent 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 is above the global average quality of users in the network ($\mu = 0.26$, as per Fig. 5 (left)), and compare it with the actual portion of users in the overall population with beauty above the global average.[4] We find that *A2: the majority illusion holds in our data sample. Overall, 43 percent of the users typically produce content with above-average quality; however, 65 percent of the population has more than 43 percent of their friends with above-average quality.* The phenomenon is very strong for the nearly 20 percent of users who have more than 86 percent 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* [60], 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 $\bar{b} = k$, for all possible values of user beauty

$$b_{nn}(k) = \frac{1}{|\{i : \bar{b}(i) = k\}|} \cdot \sum_{i:\bar{b}(i)=k} \frac{\sum_{j \in \Gamma_{out}(i)} \bar{b}(j)}{|\Gamma_{out}(i)|}, \qquad (2)$$

where $\bar{b}(i)$ is user i's beauty and $\Gamma_{out}(i)$ are i's out-neighbors. Fig. 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 $\rho = 0.48$) reveals an assortative trend,

4. 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.

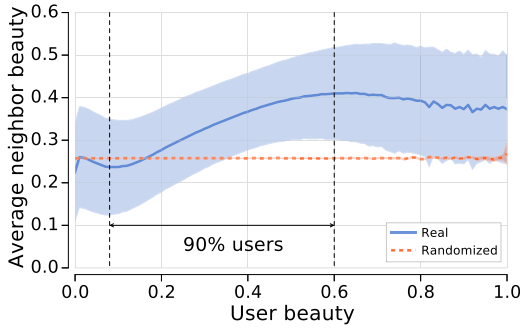

Fig. 7. Correlation spectrum of user beauty on the Flickr follower network. The highlighted interval on the beauty axis includes the user beauty values of 90 percent of the population. Variance is shown. The correlation spectrum for a null-model with randomly reshuffles user beauty scores is reported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

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 percent of our sample's population. To tell apart real any assortativity trend from statistical artifacts, we need to compare the results obtained on the real data 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 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 [61]. 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 [62], [63] 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 [64], content quality has never been investigated in relation to such network properties. We aim to shed light on this relationship by answering two research questions that help explain the assortative trend we found.

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 ebauty 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 [65], 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.

6.2.2 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 [31], [66] 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 [32], [67], [68] and to measure the equivalence between treatment and control groups. Here we borrow a framework introduced by Rubin [69] and later summarized by Stuart [32], which has been successfully used in other observational studies aimed at infer causality [33]. 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 (σ) in the treated group, is under a given threshold commonly set to 0.25. Formally

$$SB_X(G_t, G_c) = \frac{\bar{X}_t - \bar{X}_c}{\sigma(X_t)} \leq 0.25. \quad (3)$$

TABLE 1
Covariates Accounted for in the Matching Experiments

Category	Covariate	SB
User	Indegree	+0.16
	Outdegree	+0.21
	Number of photos uploaded	−0.16
	Number of group memberships	+0.21
	Number of favorites given	+0.18
	Number of favorites received	+0.17
	Average photo beauty	−0.07
	Weeks elapsed from join date	+0.19
Neighbors	Number of photos uploaded	+0.18
	Average photo beauty	+0.22
New neighbors	Number of photos uploaded	+0.18

The variables considered are measured for three types of users: i) *the users who create new links,* ii) *their neighbors before the action of link creation, and* iii) *their new neighbors. All the measurements are taken in the week of link creation. The standardized bias values (SB) for the first matching experiment are reported.*

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 percent 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 $SB(G_t, G_c) \leq 0.25$ 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 2.7 M 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 [32]: 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.

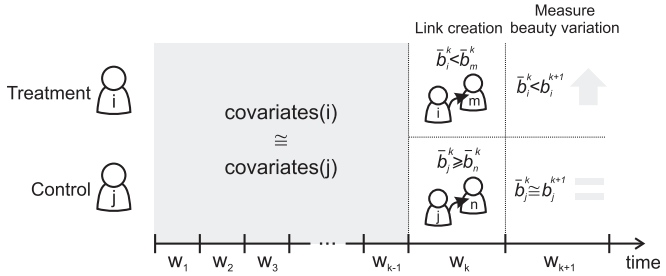

Fig. 8. Simplified example of matching experiment setup, with just one user in each group. The two users are statistically equivalent with respect to all the considered covariates measured at week $k-1$. At week k, user i creates a link towards user m, whose photos posted until week k have higher beauty than i's ($\bar{b}^k(i) < \bar{b}^k(m)$). User j instead, creates a new link towards user n, whose beauty is not higher than his ($\bar{b}^k(j) \geq \bar{b}^k(n)$). User i is the treatment user, user j is the control one. At week $k+1$, both users will post new photos; the hypothesis is that the i's new photos will have higher quality than i's previous quality ($b^{k+1}(i) > \bar{b}^k(i)$), whereas no statistically significant variation will occur to j's beauty.

After the two groups are built, we execute the algorithm described in the previous section (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 Δ_b 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 the group, the outcome variable is defined as follows:

$$\Delta_b(G_t) = \frac{1}{|G_t|} \cdot \sum_{i \in G_t} \frac{b^{w+1}(i)}{\bar{b}^w(i)}; (\text{same for } G_c). \tag{4}$$

Fig. 8 depicts a simplified sketch of the matching experiment.

The measure of Δ_b confirms our hypothesis: the treatment group experiences an average 2 percent increase in Δ_b, whereas no significant increase is found in the control group (Fig. 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

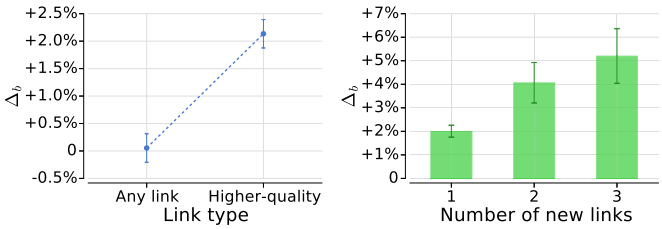

Fig. 9. Matching experiment. Beauty increase Δ_b after a generic link creation (control) versus a link creation towards a user with higher beauty (treatment). Beauty increase after the creation of n links towards users with higher beauty. Ninety-five percent confidence intervals are shown.

includes in G_t only users who created *exactly* $n \in \{1, 2, 3\}$ links towards higher-quality users. We limit ourselves to $n = 3$ because for larger n we could not form matching pairs of treatment and control groups large enough to ensure statistical significance. We find that the influence effect accumulates with new connections, with diminishing returns (Fig. 9, right). Last, to measure how much the beauty increase depends on the *magnitude* of the difference between the user's beauty and that of its new neighbors, we restrict G_t to the users whose new neighbors at week w ($\Gamma^w(i)$) have an average beauty that is α times greater than their own

$$\bar{b}^w(\Gamma^w(i)) = \frac{1}{|\Gamma^w(i)|} \cdot \sum_{j \in \Gamma^w(i)} \bar{b}^w(j), \tag{5}$$

$$\bar{b}^w(\Gamma^w(i)) \geq (1 + \alpha) \cdot \bar{b}^w(i).$$

We find that, the greater the beauty differential, the greater the increase—noticeable until $\alpha = 0.5$, after which the confidence interval becomes too wide to make any assessment (Fig. 10).

In summary, we found that *A4: users' produced quality increases as a result of new established connections with higher-quality users; the higher the number of those new contacts and the higher their quality, the stronger the effect*.

Beauty Imbalance Kills. Finally, to answer question *Q5*, we set up an experiment to ascertain if strong quality imbalance reduces user engagement. Also for this experiment we use a weekly-quantized timeline, but this time we partition users among G_c and G_t based on their existing neighbor set rather than on the new connections they create. For every week-user pair (w, i) we measure the average beauty of i's full neighbor set at week w, namely $\bar{b}^w(\Gamma^w(i))$ as defined in Equation (6). We measure how much the average neighbor beauty deviates from the user beauty

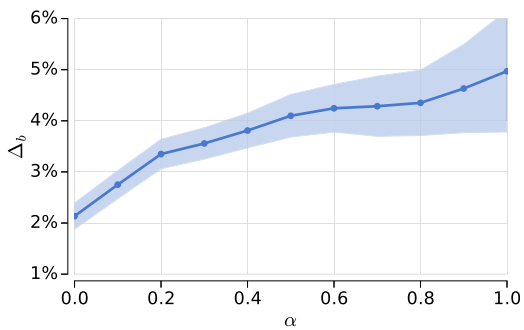

Fig. 10. Beauty increase of a user after creation of links towards users with quality α times higher. Ninety-five percent confidence interval is shown.

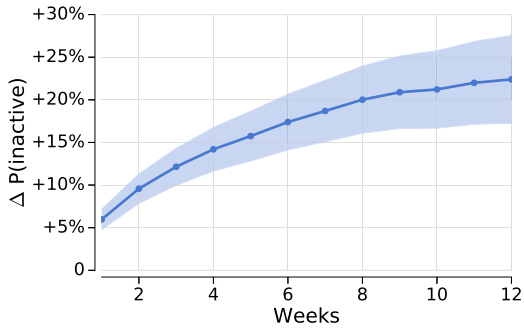

Fig. 11. Increase in the probability of becoming inactive for n weeks for users with high beauty imbalance with their neighbors, compared to balanced users. Ninety-five percent confidence interval is shown.

$$\bar{b}^w(i) + \delta \cdot \bar{b}^w(i) = \bar{b}^w(\Gamma^w(i)). \tag{6}$$

If the two quantities are in the same close range ($-0.1 \leq \delta \leq 0.1$), we add the user to G_c. Else, if the difference is substantial—namely 30 percent or more ($\delta \geq 0.3$)—we add it to G_t. We then measure the proportion p^{inact} of users in each group who remain inactive (i.e., no photo uploads) for n consecutive weeks after week w and compute the ratio between the values for the two sets (p_t^{inact}/p_c^{inact}) to measure the relative increment in treatment over control. We observe that the treatment group has higher probability of inactivity that grows from $+5$ to $+20$ percent in the first 12 weeks (Fig. 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 [70], thus increasing the linkage 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 the gap statistic [71]. 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

TABLE 2
Clustering Results

	%users	beauty	fav/photo	connects
Low quality	41.2%	0.17	0.00	0.06
Forlorn beauty	28.1%	0.42	0.01	0.10
Regular user	22.1%	0.25	0.01	0.21
Superstar	8.6%	0.42	0.15	0.35

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.

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.[5]

The clustering results confirm that the talent of a large portion of the user-base—more that $1/4$th 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 worsen 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.

5. https://www.flickr.com/explore

TABLE 3
Average Value of Descriptive Metrics for Users in Different Clusters

	Low	*Forlorn*	*Regular*	*Superstar*
photo count	1,060	200.4	1869	822.4
time on platform	104.4	84.68	187.0	198.3

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 ($BB_{\pm 10}$) that is within a small range from the user beauty of the recommendation recipient (± 10 percent), in order to avoid quality imbalance. We simulate both approaches on a random sample of 400 K 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 $b_{recs} = \frac{1}{|R|}\sum_{(u,r)\in R} \bar{b}(r)$.
- Average ratio between the user beauty of the recommendation recipient and the user beauty of the recommended contact $b_{ratio} = \frac{1}{|R|}\sum_{(u,r)\in R} \frac{\bar{b}(u)}{\bar{b}(r)}$; 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 $fav_{recs} = \frac{1}{|R|}\sum_{(u,r)\in R} fav(r)$.
- Oortion of users in the *Forlorn Beauty* cluster in the recommendation list $p_{forlorn} = \frac{\sum_{(u,r)\in R} \mathbf{I}(r \in ForlornBeautySet)}{|R|}$, where $\mathbf{I}(\bullet) = 1$ if the condition of its argument is true, 0 otherwise.

Following the findings from Section 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. Fig. 12 shows the results of the two approaches.

The CN approach selects contacts with beauty higher than $BB_{\pm 10}$, but only slightly higher, considering the strict ± 10 percent constraint in $BB_{\pm 10}$, which, by definition, will limit the maximum level of beauty for new contacts of a

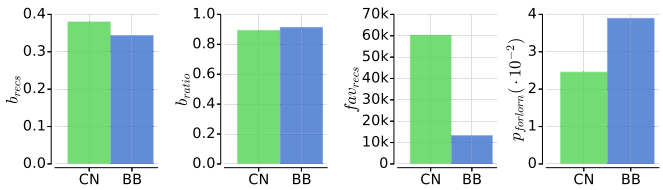

Fig. 12. Comparison between a friend-of-friend approach based on common neighbors (CN) and a quality-oriented algorithm that suggests users at distance 2 with the highest average beauty score ($BB_{\pm 10}$) in a ± 10 percent quality interval.

given node. On the other hand, CN's recommended users are 5 times more popular in terms of number of favorites. This introduces higher beauty imbalance ($+2.35$ percent) and does not reach *Forlorn* users as effectively: $BB_{\pm 10}$ suggests 49 percent 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 [72]. 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 versus 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 the 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 percent 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

[1] M. J. Halvey and M. T. Keane, "Exploring social dynamics in online media sharing," in *Proc. Int. Conf. World Wide Web*, 2007, pp. 1273–1274. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242804

[2] A. Susarla, J.-H. Oh, and Y. Tan, "Social networks and the diffusion of user-generated content: Evidence from YouTube," *Inf. Syst. Res.*, vol. 23, no. 1, pp. 23–41, 2012.

[3] A. Bouch, A. Kuchinsky, and N. Bhatti, "Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2000. [Online]. Available: http://doi.acm.org/10.1145/332040.332447

[4] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman, "Determining causes and severity of end-user frustration," *Int. J. Human-Comput. Interaction*, vol. 17, no. 3, pp. 333–356, 2004.

[5] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 4, pp. 241–257, 2006.

[6] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 457–466. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654927

[7] X. Jin, J. Chi, S. Peng, Y. Tian, C. Ye, and X. Li, "Deep image aesthetics classification using inception modules and fine-tuning connected layer," in *Proc. 8th Int. Conf. Wireless Commun. Signal Process.*, 2016, pp. 1–6.

[8] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–679.

[9] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 497–506.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis.*, 2006. [Online]. Available: http://dx.doi.org/10.1007/11744078_23

[11] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 419–426.

[12] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 33–40.

[13] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1657–1664. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995467

[14] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1784–1791.

[15] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2408–2415.

[16] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 386–399.

[17] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver, "The role of tags and image aesthetics in social image search," in *Proc. Int. Conf. Weblogs Social Media*, 2009. [Online]. Available: http://doi.acm.org/10.1145/1631144.1631158

[18] O. Wu, W. Hu, and J. Gao, "Learning to predict the perceived visual quality of photos," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 225–232.

[19] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

[20] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 145–152.

[21] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes, "6 seconds of sound and vision: Creativity in micro-videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4272–4279.

[22] M. Redi and B. Merialdo, "Where is the beauty?: Retrieving appealing VideoScenes by learning Flickr-based graded judgments," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 1363–1364. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396486

[23] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1113–1119.

[24] O. Russakovsky, et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[25] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *Int. J. Human-Comput. Studies*, vol. 60, no. 3, pp. 269–298, 2004.

[26] A. De Angeli, A. Sutcliffe, and J. Hartmann, "Interaction, usability and aesthetics: What influences users' preferences?" in *Proc. 6th Conf. Designing Interactive Syst.*, 2006, pp. 271–280.

[27] K. Zhou, M. Redi, A. Haines, and M. Lalmas, "Predicting pre-click quality for native advertisements," in *Proc. Int. Conf. World Wide Web*, 2016, pp. 299–310. [Online]. Available: https://doi.org/10.1145/2872427.2883053

[28] Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes, "To click or not to click: Automatic selection of beautiful thumbnails from videos," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 659–668. [Online]. Available: http://doi.acm.org/10.1145/2983323.2983349

[29] R. Schifanella, M. Redi, and L. M. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2015, pp. 397–406.

[30] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proc. Int. Conf. World Wide Web*, 2012. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187907

[31] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc. Nat. Academy Sci. United States America*, vol. 106, no. 51, pp. 21544–21549, 2009.

[32] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statist. Sci.*, vol. 25, no. 1, pp. 1–21, Feb. 2010.

[33] T. Althoff, P. Jindal, and J. Leskovec, "Online actions with offline impact: How online social networks influence online and offline user behavior," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2017, pp. 537–546. [Online]. Available: http://doi.acm.org/10.1145/3018661.3018672

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[35] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Waltham, MA, USA: Focal Press, 2007.

[36] B. Hurter, *Portrait Photographer's Handbook*. Buffalo, NY, USA: Amherst Media, Inc, 2007.

[37] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2206–2213.

[38] P. Obrador, M. A. Saad, P. Suryanarayan, and N. Oliver, *Towards Category-Based Aesthetic Models of Photographs*. Berlin, Germany: Springer, 2012.

[39] M. Redi, N. Rasiwasia, G. Aggarwal, and A. Jaimes, "The beauty of capturing faces: Rating the quality of digital portraits," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2015, pp. 1–8.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[41] B. Thomee, et al., "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[42] W. Mason and S. Suri, "Conducting behavioral research on Amazon's mechanical turk," *Behavior Res. Methods*, vol. 44, no. 1, pp. 1–23, 2012. [Online]. Available: http://dx.doi.org/10.3758/s13428–011-0124-6

[43] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal," in *Proc. 6th Int. Workshop Quality Multimedia Experience*, 2014, pp. 29–34.

[44] Y. Fu, T. Hospedales, T. Xiang, S. Gong, and Y. Yao, "Interestingness prediction by robust learning to rank," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 488–503. [Online]. Available: http://dx.doi.org/10.1007/978–3-319-10605-2_32

[45] E. Siahaan, J. A. Redi, and A. Hanjalic, "Beauty is in the scale of the beholder: A comparison of methodologies for the subjective assessment of image aesthetic appeal," in *Proc. 2nd ACM Int. Workshop Crowdsourcing Multimedia*, 2013, pp. 29–34.

[46] W.-T. Chu, Y.-K. Chen, and K.-T. Chen, "Size does matter: How image size affects aesthetic perception?" in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 53–62. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502102
[47] M. A. Hagen and R. K. Jones, "Cultural effects on pictorial perception: How many words is one picture really worth?" in *Perception and Experience*, R. Walk, J. Pick, and L. Herbert, Eds. New York, NY, USA: Springer, 1978, pp. 171–212. [Online]. Available: http://dx.doi.org/10.1007/978–1-4684-2619-9_6
[48] J. A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, pp. 102–141, 1994.
[49] Y. Miyamoto, R. E. Nisbett, and T. Masuda, "Culture and the physical environment: Holistic versus analytic perceptual affordances," *Psychological Sci.*, vol. 17, no. 2, pp. 113–119, 2006.
[50] K. Dewar, W. M. Li, and C. H. Davis, "Photographic images, culture, and perception in tourism advertising," *J. Travel Tourism Marketing*, vol. 22, no. 2, pp. 35–44, 2007.
[51] K. Yanai and B. Qiu, "Mining cultural differences from a large number of geotagged photos," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 1173–1174. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526914
[52] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *British Med. J.*, vol. 314, no. 7080, 1997, Art. no. 572.
[53] J. Ye, J. Li, M. G. Newman, R. B. Adams, and J. Z. Wang, "Probabilistic multigraph modeling for improving the quality of crowdsourced affective data," *IEEE Trans. Affective Comput.*, to be published. doi: 10.1109/TAFFC.2017.2678472.
[54] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Appl. Statist.*, vol. 28, pp. 20–28, 1979.
[55] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. Conf. Internet Meas. Conf.*, 2007. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298311
[56] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies: Social link prediction from shared metadata," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2010. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718521
[57] S. L. Feld, "Why your friends have more friends than you do," *Amer. J. Sociol.*, vol. 96, pp. 1464–1477, 1991.
[58] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 225–233.
[59] K. Lerman, X. Yan, and X.-Z. Wu, "The majority illusion in social networks," *PloS One*, vol. 11, no. 2, 2016, Art. no. e0147617.
[60] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[61] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," *ACM Trans. Web*, vol. 6, no. 2, Jun. 2012, Art. no. 9. [Online]. Available: http://doi.acm.org/10.1145/2180861.2180866
[62] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401897
[63] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, "Feedback effects between similarity and social influence in online communities," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401914
[64] F. Dobrian, et al., "Understanding the impact of video quality on user engagement," in *Proc. Conf. Appl. Technol. Archit. Protocols Comput. Commun.*, 2011. [Online]. Available: http://doi.acm.org/10.1145/2018436.2018478
[65] P. Blau, *Exchange and Power in Social Life*. Hoboken, NJ, USA: Wiley, 1964. [Online]. Available: https://books.google.it/books?id=qhOMLscX-ZYC
[66] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociol. Methods Res.*, vol. 40, no. 2, pp. 211–239, 2011.
[67] P. R. Rosenbaum, "Observational studies," in *Observational Studies*. Berlin, Germany: Springer, 2002, pp. 1–17.
[68] A. Olteanu, O. Varol, and E. Kiciman, "Distilling the outcomes of personal experiences: A propensity-scored analysis of social media," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2017, pp. 370–386. [Online]. Available: http://doi.acm.org/10.1145/2998181.2998353
[69] D. B. Rubin, "Using propensity scores to help design observational studies: Application to the tobacco litigation," *Health Serv. Outcomes Res. Methodology*, vol. 2, no. 3, pp. 169–188, 2001.
[70] J. Su, A. Sharma, and S. Goel, "The effect of recommendations on network structure," in *Proc. Int. Conf. World Wide Web*, 2016. [Online]. Available: http://dx.doi.org/10.1145/2872427.2883040
[71] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Statist. Soc.*, vol. 63, no. 2, 2001. [Online]. Available: http://dx.doi.org/10.1111/1467–9868.00293
[72] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. Int. AAAI Conf. Weblogs Social*, 2010, pp. 11–13.

Luca Maria Aiello received the PhD degree in computer science from the University of Torino, in 2012. He is senior research scientist at Nokia Bell Labs and research fellow at the ISI Foundation in Turin, Italy. He has been a research scientist in the Yahoo Labs for almost 5 years. He conducts interdisciplinary research in network science, computational social science, and urban informatics. He co-authored 50+ peer-reviewed papers and his work has been covered by more than 200 news articles published by prestigious news outlets worldwide including Wired, *Wall Street Journal*, and BBC. He is a founding member of GoodCityLife.org, a global network of scientist with the goal of giving a good life to city dwellers.

Rossano Schifanella is an assistant professor in computer science with the University of Turin, a visiting scientist at Nokia Bell Labs, and a former visiting scientist at the Yahoo Labs and with Indiana University School of Informatics and Computing. His research embraces the creative energy of a range of disciplines across technology, computational social science, data visualization, and urban informatics. He is interested in computational methods to investigate social phenomena, aesthetics and creativity in media platforms, and figurative language. He is also passionate about building mapping tools that capture the sensorial layers of a city and designing innovative methods to explore urban spaces.

Miriam Redi received the PhD degree from the Multimedia group in EURECOM, Sophia Antipolis. She is a research scientist on the Social Dynamics team at Bell Labs Cambridge, where her research focuses on content-based social multimedia understanding and culture analytics. In particular, she explores ways to automatically assess visual aesthetics, sentiment, and creativity and exploit the power of computer vision in the context of web, social media, and online communities. Previously, she was a postdoc in the Social Media group, Yahoo Labs Barcelona and a research scientist in Yahoo London.

Stacey Svetlichnaya received the BS and MS degrees in symbolic systems from Stanford University, in 2011 and 2012, respectively. She is a research engineer on the Yahoo Computer Vision & Machine Learning team in San Francisco, California. Her recent deep learning research includes object recognition, image aesthetic quality and style classification, photo caption generation, and modeling emoji usage. She has worked extensively on Flickr image search and data pipelines, as well as automating content discovery and recommendation. Prior to Flickr, she helped develop a visual similarity search engine with LookFlow, which Yahoo acquired in 2013.

Frank Liu received the BS (honors) degree from Stanford University and the MS degree in electrical engineering from Stanford University. He is a software developer in the Flickr Vision team. During that time, he completed a minor in computer science. At Flickr, he works on scalable training and deployment of deep learning models for object classification, saliency, aesthetics, OCR, and other computer vision applications.

Simon Osindero is a pioneer in the field of machine learning and was the co-inventor of deep belief networks whilst researching as a post-doctoral fellow in the Hinton Group, University of Toronto. In his current role as an A.I. architect with Yahoo, he leads computer vision and machine learning R&D at Flickr. He joined Yahoo in 2013 after it acquired LookFlow, a company he co-founded in 2009 to productize cutting edge research from the fields of machine learning and human-computer interaction. Prior to starting LookFlow, he worked with a Montreal-based start-up, Idilia, designing machine-learning algorithms for natural language processing.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.