A meritocratic network formation model for the rise of social media influencers

Nicolò Pagan (pagann@ethz.ch)
Swiss Federal Institute of Technology in Zurich
https://orcid.org/0000-0003-0071-7856

Wenjun Mei
ETH Zurich

Cheng Li
Swiss Federal Institute of Technology in Zurich

Florian Dörfler
Swiss Federal Institute of Technology in Zurich

Article

Keywords: User-Generated Content (UGC), social media, meritocratic network formation model

Posted Date: April 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-424956/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at Nature Communications on November 30th, 2021. See the published version at https://doi.org/10.1038/s41467-021-27089-8.
A meritocratic network formation model for the rise of social media influencers

Nicolò Pagan, Wenjun Mei, Cheng Li, and Florian Dörfler

Abstract

Many of today’s most used online social networks such as Instagram, Youtube, Twitter, or Twitch are based on User-Generated Content (UGC), and the exploration of this content is enhanced by the integrated search engines. Prior multidisciplinary effort on studying social network formation processes has privileged topological elements or socio-strategic incentives. Here, we propose an untouched meritocratic approach inspired by empirical evidence on Twitter data: actors continuously search for the best UGC provider. We statistically and numerically analyze the network equilibria properties: while the expected outdegree of the nodes remains bounded by the logarithm of the network size, the expected indegree follows a Zipf’s law with respect to the quality ranking. Notably, our quality-based mechanism provides an intuitive explanation of the origin of the Zipf’s regularity in growing social networks. Our theoretical results are empirically validated against large data-sets collected from Twitch, a fast-growing platform for online gamers.

Introduction

Especially since the explosion of online services in the past couple of decades, the impact of social networks on our lives has become more and more multifaceted: they play central roles, e.g., in the dissemination of information [1], in the adoption of new technologies [2], in the diffusion of healthy behavior [3], in the formation and polarization of public opinion [4, 5]. To advance our understanding on the phenomena that take place within these platforms, there has been a recent coming-together of multiple disciplines in the study of social networks. Much of the attention has been devoted on measuring the macroscopic properties of these networks, e.g., degree, density, or connectivity, as well as on understanding the microscopic formation mechanisms [6].

One of the most noticeable distinguishing features of some of today’s most popular online platforms such as Twitter, Instagram, YouTube, or TikTok, is that, unlike others, e.g., Facebook or LinkedIn, they are directed networks, hence users can follow real-life strangers, e.g., famous people. Another crucial characteristic is that they allow people to contribute with their user-submitted tweets, photos, text posts, or videos, to the point that this User-Generated Content (UGC) has become the lifeblood of these online social networks [7, 8]: in 2020, every day 500 million tweets were sent [9], and more than 80 million Instagram pictures were posted [10]. Thanks to the use of hashtags and the integrated search engines, this UGC can be easily explored by platforms’ users in search of new content or connections. As a result, these directed online social networks do not necessarily mimic real-life relationships. Rather, they witness the interaction between users who eventually share interests, but are otherwise strangers [11].

In social media marketing, UGC is considered an effective form of word-of-mouth publicity [12, 13]. Compared to traditional brand-generated content, UGC gives a genuine feeling, hence it enhances users’ trust [14], and ultimately it has a stronger impact on consumers’ purchase behavior [15]. This opportunity has triggered the rise of the “new” influencers [16], e.g., fashion enthusiasts who rapidly gain popularity on Instagram thanks to the platform’s capability to reach wide audiences [17], way beyond real-life friends. As a result, in 2017 more than 70% of US businesses engaged Instagram influencers to promote their products [18].

Given the predominance of these services in our daily lives and their universal diffusion (especially within the young generations [19]), the large impact on social media marketing strategies, and the spreading potential of these highly influential nodes [20], it is important to understand (i) how the UGC relates with the platform’s capability to reach wide audiences [17], way beyond real-life friends. As a result, in 2017 more than 70% of US businesses engaged Instagram influencers to promote their products [17].

Stochastic Actor-Oriented Models (SAOM) [25], in sociology, and strategic network formation models [26], in economics, assume actors decide their ties according to a utility maximization principle based on sociological elements, such as reciprocity or network closure [27], or topological measures, e.g., indegree or closeness centrality [28], or a combination of them [29]. These elements are essential for real-life connections, and to a less extent for Facebook’s friendships. Yet, in nowadays’ most popular UGC-based directed social networks users are likely to interact with others they do not know in real life, thus they are not interested in reciprocity or network closure as a form of bonding social capital [11, 30]. As a matter of fact, on Instagram only 14% of the relationships are reciprocated and the average clustering coef-
ficient is smaller than 10%, while on Facebook reciprocity and clustering are found to be respectively 100% (the network being undirected) and 30% [31]. Moreover, since these platforms are based on directed networks, users cannot control their indegree.

Since the seminal paper on the random graph model by Erdős and Rényi [32], also the complex networks community developed a multitude of network formation models based on topological elements (see [33, 34, 35] for extensive surveys). Among them, the preferential attachment model, proposed by Barabási and Albert [36], in which newborn nodes choose connections proportional to the degree, has been widely acknowledged. While this mechanism leads to the scale-free effect observed in many real-world networks [37], this “rich get richer” philosophy does not justify the rise of “new” Instagram celebrities, i.e., the so-called Instafamous [38], whose success is built without prior fame.

The prevalence of UGC-based social media platforms and the absence of proper mathematical models inspire us to think about the network formation processes from an untouched perspective. In our proposed framework users are associated with an attribute defining the quality of their UGC which represents the likelihood the content will be liked by others. Following a stochastic process similar to the random graph model [32], agents meet with uniform probability, and strategically create connections (directed ties) according to a utility maximization principle. Notably, the individual decision is based on the UGC quality, and it is rooted in an intuitive meritocratic principle. To support this intuition, we collected and analyzed a longitudinal Twitter data-set on network scientists (which is publicly available at [39]).

We statistically study the proposed network formation dynamics as well as the network properties at equilibrium. We discover that the indegree distribution exhibits a specific pattern: the highest quality node expects to have twice (three times) as many followers as the second (third) highest, and so on. This regularity, called Zipf’s law, has been first reported in the analysis of word frequencies [40], but it has been also found in many systems, including growing social networks [41]. Conversely, the outdegree distribution is similar to a log-normal distribution, with expectation equal to the harmonic number of the network size. Finally, the significant overlap in the followers’ sets is compatible with the homophily that characterizes agents with similar interests.

We emphasize that, despite being widely assumed to be ubiquitous for systems where objects grow in size or are fractured through competition [42], the principle underlying the origin of Zipf’s law is an open research question (see [43] for a survey). Thus, a fundamental contribution lies in our quality-based rule revealing an intuitive, meritocratic mechanism for the emergence of Zipf’s law. Furthermore, although an underlying power-law distribution is necessary, “there is more than a power-law in Zipf” [42]. Thus, our model unfolds this regularity among network influencers, while preserving the well-known scale-free effect.

Finally, in order to empirically validate our model, we collected three data-sets (which are publicly available [44]) from Twitch, a fast-growing platform for online gamers [45]. The successful comparison of our theoretical predictions against them indicates that our model, despite being of simple and parsimonious form, already captures several real-world network properties.

Results

The majority of today’s online social networks offers the users the possibility to actively contribute to the platform’s growth by sharing different forms of UGC, according to their interests, competences, and willingness. Looking from a different perspective, users are also exposed to the content generated by others in the platform. Thanks to the integrated search engines and the use of hashtags they can explore this content, discover users with similar interests, and ultimately decide to become followers. Given the limited time they can spend on social media platforms, users seek to optimize the list of folloowees so to receive content they are really interested in. Intuitively, the decision to follow other accounts roots in a meritocratic principle based on UGC’s quality.

Meritocratic Principle

In order to support our intuition, we collected a longitudinal Twitter data-set [39] on a network composed of more than 6000 scientists working in the area of complex social networks. Compared to other data-sets, one of the advantages is that we can expect most of the complex network scientists to be active on Twitter, given that they consistently study social networks effects. Moreover, the most popular nodes can be easily associated to renowned researchers in the field. Arguably, the number of followers can be considered as a proxy for the quality of the content generated by a user. We confirm this hypothesis by manually inspecting and labelling the top-indegree nodes.

Then, we enumerate the agents in decreasing-indegree order, so that agent 1 (presumably providing the best UGC) is the most followed, agent 2 is the second most, and so on. For each agent $i$, we reconstruct the temporal sequence of the outgoing connections $(i_1, i_2, \ldots, i_{d_{i}^{\text{out}}})$, where $i_k$ is the index (rank) of the destination of the $k$-th outgoing connection of agent $i$, and $d_{i}^{\text{out}}$ is her final outdegree. Then, for each agent $i$ we compute the probability

$$p_i = \frac{\left| \left\{ j \in \{2, \ldots, d_{i}^{\text{out}}\} : \text{s.t. } i_j > \text{median} \{i_1, \ldots, i_{j-1}\} \right\} \right|}{d_{i}^{\text{out}} - 1},$$

that estimates the likelihood that a new connection of agent $i$ is higher (in ranking) than the median of the previous connections. At the extreme cases, for $p_i = 0$ the sequence $(i_1, i_2, \ldots, i_{d_{i}^{\text{out}}})$ is such that every new folloowee has rank smaller than (or equal to) the median of the current list of folloowees’ ranks, i.e., $i_j \leq \text{median} \{i_1, \ldots, i_{j-1}\}$. As such, the rolling median, i.e., the median computed among the first $k$ elements, is always non-increasing in $k$, and the user is continuously seeking better UGC. Conversely, for $p_i = 1$ the rolling median is always increasing.

The histogram of the distribution of $p_i$ is shown in Fig. 1, in purple. We compare this distribution with what we call the null hypothesis, plotted in blue, in which we remove the temporal-ordered pattern of the sequence. The elements in agent $i$’s folloowees set $\mathcal{F}_{i}^{\text{out}} := \{i_1, i_2, \ldots, i_{d_{i}^{\text{out}}}\}$ are randomly re-ordered: in the new sequence $(i_1, i_2, \ldots, i_{d_{i}^{\text{out}}})$, the $k$-th element of the list is drawn uniformly at random from $\mathcal{F}_{i}^{\text{out}} \setminus \{i_1, \ldots, i_{k-1}\}$. The null hypothesis has a median value of approximately 0.5, which is easy to interpret: if the sequence is completely random, adding
weighted directed network among \( N \) actors. To formalize our quality-based model, we consider the unweighted directed network among \( N \) actors.

The distribution is 0.42. In blue, we compute the same distribution upon reshuffling the temporal sequences of the connections (null hypothesis). The median of this distribution is 0.49. The two distributions are statistically significantly different (p-value of Kolmogorov-Smirnov test \( \ll 0.001 \)).

![Figure 1](image)

**Figure 1.** Results from the Twitter data-set. In purple, we plot the histogram of the probability \( p_i \) as defined in (1). The median of the distribution is 0.42. In blue, we compute the same distribution upon reshuffling the temporal sequences of the connections (null hypothesis). The median of this distribution is 0.49. The two distributions are statistically significantly different (p-value of Kolmogorov-Smirnov test \( \ll 0.001 \)).

an extra element to the partial sequence has 50% probability of being above the median, and 50% probability of being below it. Comparing the two distributions, we notice that the median of the data is smaller than the one computed for the null hypothesis, indicating that agents tend to have a decreasing ranking sequence of followees (corresponding to an increasing ranking sequence of their quality). Since the difference is statistically significant, we can reject the hypothesis that the temporal sequence is random.

Ultimately, this empirical evidence confirms our intuition, i.e., users tend to continuously increase the quality threshold of the new followees. This characteristic, though, is typically missing in the network formation literature, as there is no quality associated to the users. For instance, in the preferential attachment model \([36]\), each user selects \( m \) followees proportional to their indegree. If the network is large enough, the probability of selecting a node \( k \) in the \( d \)-th draw does not depend on \( d \). Therefore, the temporal sequence of connections of the preferential attachment model is similar to the null-hypothesis.

**Quality-based model**

To formalize our quality-based model, we consider the unweighted directed network among \( N \geq 2 \) agents whose UGC revolves around a specific common interest, e.g., a particular travelling destination. We denote the directed tie from \( i \) to \( j \) with \( a_{ij} \in \{0, 1\} \), where \( a_{ij} = 1 \) means \( i \) follows \( j \). Then, we assume there are no self-loops and that each agent \( i \) can only control her followees \( a_{ij} \) but not her followers \( a_{ji} \). Similarly to the approach in \([24]\), we endow each actor \( i \) with an attribute \( q_i \), drawn from the uniform probability distribution in \([0, 1]\), that describes the average quality of \( i \)'s content, e.g., a picture taken at that travelling destination. The quality \( q_i \) can be seen as the expectation of a Bernoulli random variable \( Q_i \) describing the probability of followers liking agent \( i \)'s content. Higher values of \( q_i \) are then associated with better UGC.

We then consider a sequential dynamical process starting from the empty network, where at each time-step \( t \in \{1, 2, \ldots \} \) each actor \( i \) picks uniformly at random another distinct actor \( j \). To reflect the meritocratic principle, we base the decision on the tie formation on the comparison between \( i \)'s current followees’ and \( j \)'s qualities. Let the payoff function of agent \( i \) measure the maximum quality received by \( i \), i.e.,

\[
V_i(t) := \max_{j \in \mathcal{F}_{i}(t)} q_j,
\]

where \( \mathcal{F}_{i}(t) := \{ j, \text{ s.t. } a_{ij}(t) = 1 \} \) denotes the set of \( i \)'s followees at time \( t \). According to a utility maximization principle, we define the update process through the following rule:

\[
a_{ij}(t+1) = \begin{cases} 
1, & \text{if } q_j > V_i(t) \\
 a_{ij}(t), & \text{otherwise},
\end{cases}
\]

meaning that \( i \) will add \( j \) in her followees’ set if \( j \) provides better quality content compared to \( i \)'s current followees.

Note that, if \( i \) finds a node that already belongs to her set of followees, the connection will not be re-discussed. While, intuitively, this may lead to a large outdegree, we will show that this is not the case. In fact, the quality threshold reflects a meritocratic principle in our payoff function: the cost of poor-quality connections is infinitely high, and the cost of good-quality ones is infinitely low.

We emphasize that the choice of the payoff function (2) reflects the natural continuous chase for the maximum quality \([46]\) (exploitation) while minimizing the effort due to non-improving connections. Alternatively, one could consider smoother payoffs such as the followees’ average quality

\[
V_i(t) := \frac{\sum_{j \in \mathcal{F}_{i}(t)} q_j}{|\mathcal{F}_{i}(t)|},
\]

as in \([24]\), which allows for more exploration, at the expense of the least-effort principle. Both definitions, though, share the same meritocratic principle previously discussed.

**Convergence**

A natural question that arises when defining a dynamical process is whether it reaches or not an equilibrium. In what follows we show that an equilibrium state is reached almost surely. In order to do so, from now on we assume that there exist no two agents with equal quality. Yet, we emphasize that the model and our analysis can be extended to the case where two or more agents have the same quality. Then, without loss of generality, we can re-order the agents by decreasing quality \( q_i \), i.e.,

\[
q_1 > q_2 > \cdots > q_N.
\]

In this way, agent 1 is the top-quality agent, agent 2 is the second best, and so on. According to our dynamics, any node \( i > 1 \) creates new links towards increasingly-quality agents, until finding
the top-quality node 1. Likewise, node 1 creates new links until finding the second highest quality agent, node 2. Convergence is guaranteed by the following theorem, proven in the Methods section.

Theorem (Convergence). For any set of qualities \( \{q_1, \ldots, q_N\} \), the network reaches an equilibrium almost surely.

An example of the resulting network is shown in Fig. 2 (darker nodes are associated with higher quality attribute).

**Figure 2.** The plot shows a network of 201 nodes upon reaching the equilibrium of the quality-based model dynamics. Nodes’ color is proportional to their quality attribute \( q_i \in [0,1] \), with darker nodes having higher quality. The size of the nodes, instead, is proportional to their indegree at equilibrium. The correlation between quality (color) and indegree (size) of the nodes is a consequence of the meritocratic principle.

**Indegree distribution**

The network structure at equilibrium is a consequence of the random selection process. Even though there can be multiple equilibria, some network macroscopic properties can be statistically described. An obvious observation is that, at equilibrium, every other node follows node 1, and node 1 follows node 2. Before studying the stationary state, we statistically describe the transient indegree distribution as a function of the quality ranking \( i \). In particular, we provide an analytical formula for the expected indegree of each node as a function of \( i \), as stated in the following theorem.

**Theorem** (Indegree distribution). The probability that node \( i \) is followed by node \( j \neq i \) after \( t > 0 \) time-steps is:

\[
P[a_{ji}(t) = 1] = \begin{cases} 
\bar{p}_i(t) := \frac{1}{t-1} \left( 1 - \left( \frac{N-i}{N-1} \right)^t \right), & \text{if } j < i, \\
\bar{p}_j(t) := \frac{1}{t} \left( 1 - \left( \frac{N-j}{N-1} \right)^t \right), & \text{if } j > i.
\end{cases}
\] (4)

Moreover, the probability of node \( i \) having indegree \( d^{\text{in}}_i(t) = d \in [0, N-1] \) after \( t \) time-steps is given by

\[
P[d^{\text{in}}_i(t) = d] = \sum_{k=0}^{d} \binom{i-1}{k} \bar{p}_{i}^{k} (1-\bar{p}_i)^{i-1-k} \left( N-i \right)^{d-k} \left( 1 - \frac{N-i}{N} \right)^{N-i-(d-k)},
\] (5)

where we omitted the time-step dependency on \( \bar{p}_i \) and \( p_j \). Finally, the expected indegree of node \( i \) after \( t \) time-steps reads as:

\[
E[d^{\text{in}}_i(t)] = \frac{N}{t} - \left( \frac{N-i}{N-1} \right)^t + \frac{N-i}{i} \left( \frac{N-i-1}{N-1} \right)^t.
\] (6)

The proof of the theorem is provided in the Supplementary Note 1, whereas in Fig. 3a we show the probability density function, together with its expectation, for a network of 1001 agents, after \( t = 200 \) time-steps (non-stationary state). The following corollary, instead, studies the probability density functions upon reaching the network dynamics equilibrium.

**Corollary.** At equilibrium, the probability that node \( i \) is followed by node \( j \neq i \) is:

\[
P[a_{ji}(t) = 1] := \lim_{t \to \infty} P[a_{ji}(t) = 1] = \begin{cases} 
\bar{p}_i^*: := \frac{1}{t}, & \text{if } j < i, \\
\bar{p}_j^*: := \frac{1}{t}, & \text{if } j > i,
\end{cases}
\] (7)

and the expected indegree of node \( i \) reads as:

\[
E[d^{\text{in}}_i] = \begin{cases} 
\frac{N-1}{N} & \text{if } i = 1, \\
\frac{N}{t} & \text{otherwise}.
\end{cases}
\] (8)

**Proof.** First, we derive the probability of node \( i \) being followed by node \( j \) at equilibrium by taking the limit \( t \to \infty \) in (4):

\[
\lim_{t \to \infty} P[a_{ji}(t) = 1] = \begin{cases} 
\bar{p}_i^* := \frac{1}{t}, & \text{if } j < i, \\
\bar{p}_j^* := \frac{1}{t}, & \text{if } j > i.
\end{cases}
\]

Similarly, taking the limit for \( t \to \infty \) of (6) yields exactly to

\[
E[d^{\text{in}}_i] = \begin{cases} 
\frac{N-1}{N} & \text{if } i = 1, \\
\frac{N}{t} & \text{otherwise}.
\end{cases}
\]

According to the above result, at equilibrium the best content provider, node 1, receives \( N-1 \) connections, node 2 has \( N/2 \) expected followers, node 3 has \( N/3 \), and so on. The result can be intuitively reached with the following plausible reasoning: any user that has not yet found node 1 nor node 2, has the same probability of finding any of the two in the coming time-step. In expectation, in half of the cases the user will become a follower of node 2 before finding and following (necessarily) node 1. In the other half of the case, she will find node 1 before having seen node 2. Thus, the expected number of followers of node 2 is half of the expected number of followers of node 1. The reasoning can be extended for any node \( j > 1 \).
Such a regular scaling property is called Zipf’s law [40] and it is illustrated in Fig. 3b, where we plot the expected indegree of each node as a function of its quality ranking, together with the probability density functions. In log-log scale, the expected indegree perfectly matches a line with coefficient $-1$. Real world evidences of Zipf’s law have been reported in many systems, including firm sizes [47], city sizes [43], connections between web-pages [48], and also growing social networks [41]. According to this law, e.g., the biggest city in the US has a population which is roughly twice as much as the second biggest, three times as much as the third biggest, and so on [49].

To empirically validate our results in the context of online social networks, we collected and analyzed data from Twitch, an online social media platform focusing on video streaming, including broadcasts of gameplay, e-sports competitions, and real-life content. Over the past decade, Twitch gradually became one of the most popular social media platforms, reaching up to 4 million unique creators streaming each month and 17 million average daily visitors [45], and serving as virtual third place for online entertainment and social interactions [50]. Our data-sets consist of the followership networks in three different categories, i.e., poker, chess, and art, where users can live-watch broadcasters playing or discussing about these arguments. Among them, we only retain the first two categories, as the third one did not satisfy the criterion of having a baseline community of interested users. The description of our crawling method and of the collected data is provided in the Methods section and in the Supplementary Note 2. In Fig. 4 we show the indegree of the 15 most followed users in chess and poker as a function of the rank. The empirical data are fitted via linear regression (in the log-log plot) and look strikingly similar to the Zipf’s law.

\[
\text{Indegree } d_{i}^{\text{in}} \sim \frac{1}{i}
\]

**Figure 3.** Given a network of $N = 1001$ agents, the plots show the probability density functions of the indegree (derived through (5) from (4) and (7)) as a function of the rank $i$. In orange, we also plot their expected value (equations (6) and (8)). In (a), the plot refers to the theoretical results at $t = 200$, in (b), the situation upon reaching the network equilibrium. The plots are shown in log-log scale, to emphasize the Zipf’s law: $\mathbb{E}[d_{i}^{\text{in}}] = N/i$.

![Log-log plot showing empirical evidence of the Zipf’s law](image)

**Figure 4.** The log-log plot shows empirical evidences of the Zipf’s law in our Twitch data-sets. In blue, we report the data on the number of followers of the top 15 users (broadcasters) in the chess category, as well as their linear regression fit of coefficient $-1.04$ (Pearson coefficient: $-0.98$). In purple, the poker data-set shows a similar fitting coefficient of $-0.98$ (Pearson coefficient: $-0.97$). Both fits are very close to the theoretically expected Zipf’s law (of coefficient $-1$) shown in orange.

---

5/12
**Zipf’s law is more than a power-law**

The peculiarity and apparent ubiquity of Zipf’s law has triggered numerous efforts to explain its origins [43]. Despite being a discrete distribution, Zipf’s law is often associated to the continuous Pareto distribution, better known as power-law [51]. For this reason, it is frequently seen as the result of a linear preferential attachment process [36] based on Gibrat’s rule of proportional growth [52] which leads to Yule-Simon (power-law) distributions [53]. However, as noticed in [42], “there is more than a power-law in Zipf”:

“although a power-law distribution is certainly necessary to reproduce the asymptotic behavior of Zipf’s law at large values of rank \( i \), any random sampling of data does not lead to Zipf’s law and the deviations are dramatic for the largest objects”. In particular, Zipf’s law emphasizes the relation among the top-ranking elements, which essentially correspond to the most important nodes, i.e., the network influencers. The typical Zipf’s sequence \( N_i, N_i/2, N_i/3, \ldots \) can be observed only if data constitutes a “coherent” set [42]. Thus, a Zipf’s law is more insightful than just a power-law. On the other hand, the mechanism leading to a power-law does not necessarily produce a Zipf’s law.

The difference with a Pareto distribution becomes evident when considering our indegree probability density function, that can be analytically derived by using the result of the previous theorem and computing the average of each user’s probability density function

\[
P[d_i^{\text{in},*} = d] = \frac{1}{N} \sum_{i=1}^{N} P[d_i^{\text{in},*} = d].
\]

As shown in Fig. 5, the theoretical indegree probability density function follows a power-law of coefficient \( \hat{\alpha} = -2.110 \pm 0.005 \) (\( p - \text{value} < 10^{-8} \)), which is not surprising since the expected indegree is distributed according to a Zipf’s law (see again the discussion in [51]). However, notice that the power-law fit does not hold in the region of low-indegree nodes (\( d_{\text{min}} = 4 \), according to the Clauset algorithm [54]). In this region, the log-normal distribution is a better fit compared to the power-law [55]. On the other hand, the fit with a log-normal distribution is worse (compared to the power-law) in the heavy tail of our distribution, i.e., in the region of high ranking nodes. However, the most noticeable difference is that, contrary to both the power-law and the log-normal curve our theoretical distribution is not monotonically decreasing, especially in its right heavy tail. The probability of having a node of indegree \( N/2 \) is slightly larger than the probability of having a node of indegree \( N/2 \pm \alpha \), for some range of \( \alpha \geq 1 \). Moreover, the probability of having a node of indegree in a neighborhood of \( N/2 \) is larger compared to the power-law fit (chosen as base-line). Conversely, our theoretical probability of having a node of indegree slightly smaller than \( N - 1 \) is negligible (\( < 10^{-8} \)), stating that the second best node cannot be too close to the best node. This is in sharp contrast with the monotonic Pareto and log-normal distributions.

In synthesis, our probability distribution differs from Pareto and log-normal distributions when focusing on the network influencers. In this range, i.e., in the extreme right tail of the distribution, the integer effect [56] causes low accuracy of the prediction of Pareto and log-normal, which are continuous distributions. Instead, the right tail of our theoretical distribution can predict more accurately the Zipf’s regularity found on the top influencers of real-world networks, as in the Twitch data-set shown in Fig. 4.

**Outdegree distribution**

Similarly to the indegree, we can study the statistical distribution of the nodes’ outdegree at equilibrium. First, notice that, contrary to the indegree, the outdegree distribution does not depend on the rank, but it is uniform for all the nodes in the network. In fact, following the dynamics, each node creates connections to increasing quality nodes, until reaching its own equilibrium. Let \( d_N^{\text{out},*} \) be the random variable describing the outdegree (at equilibrium) of a general node \( i \) in a network of \( N \) agents. First, note that the probability of node \( i > 1 \) finding node 1 (or node 1 finding node 2) at her first choice is simply:

\[
P[d_N^{\text{out},*} = 1] = \frac{1}{N-1}.
\]

Conversely, the probability of having maximum outdegree corresponds to the probability of meeting all the other nodes in increasing quality ordering, i.e.,

\[
P[d_N^{\text{out},*} = N-1] = \frac{1}{(N-1)!}.
\]

Obviously, a node cannot have outdegree larger than \( N - 1 \) in a network of \( N \) agents, thus \( P[d_N^{\text{out},*} = d] = 0 \), for all \( d \geq N \).
shows that the expected outdegree has a similar growth with very quickly vanishing right tails.

Then, the probability \( P \) variable describing the first meeting of \( i \) values are particularly rare. From the recursive description of the nodes’ outdegree as a function of the number of nodes \( N \), it is possible to compute the expected probability density functions by means of recursion on the network size \( N \). In other words, the probability \( P \left[ d_{N^{out}} = d \right] \) for \( N \geq 2 \) and \( 1 \leq d \leq N-1 \), can be described recursively as follows:

\[
P \left[ d_{N^{out}} = d \right] = \sum_{i=1}^{d-1} P \left[ C_1 = j \right] P \left[ d_{j^{out}} = d-1 \right] + \sum_{j=1}^{N-d+1} P \left[ C_1 = j \right] P \left[ d_{j^{out}} = d-1 \right] = \frac{1}{N-1} \sum_{j=1}^{N-1} P \left[ d_{j^{out}} = d-1 \right].
\]

(9)

In other words, the probability \( P \left[ d_{N^{out}} = d \right] \) is equal to the sum of the probability of agent \( i \) choosing \( j \in \{1, \ldots, i-1, i+1, \ldots, N\} \) as first choice times the probability of having outdegree \( d-1 \) in a network with \( j \) remaining nodes (the node itself and the \( j-1 \) nodes to which she can still connect to).

Thanks to (9), we can describe the outdegree probability density functions by means of recursion on the network size \( N \). Fig. 6 illustrates it for different values of \( N \). Similarly to a log-normal distribution, it exhibits non-monotonic properties (first increasing, then decreasing) and quickly vanishing tail: extremely large values are particularly rare. From the recursive description of the probability density function it is possible to compute the expected nodes’ outdegree as a function of the number of nodes \( N \), as stated in the following theorem.

**Theorem** (Outdegree distribution). At equilibrium, the nodes’ expected outdegree \( \mathbb{E} \left[ d_{N^{out}} \right] \) in a network of \( N \geq 2 \) agents equals the \((N-1)\)-th harmonic number:

\[
\mathbb{E} \left[ d_{N^{out}} \right] = \sum_{k=1}^{N-1} \frac{1}{k}.
\]

According to the theorem, the expected outdegrees as a function of \( N \) are given by the harmonic sequence: 1, 3/2, 11/6, 25/12, … . The proof provided in the Supplementary Note 1 is built on the derivation of the outdegree probability distribution, however the result can be intuitively derived from the following observations: the expected outdegree is uniform across all the nodes, and the sum of the expected outdegree equals the sum of the expected indegree of all the nodes. From the result on the indegree distribution (see (8)), the sum of the expected indegree is equal to \((N-1) + N/2 + N/3 + \ldots + 1\). Thus, the expected outdegree is equal to \((N-1)/N + N/(2N) + N/(3N) + \ldots + 1/N\), which in fact corresponds to the \( N-1 \)-th harmonic number.

Fig. 7 shows that the expected outdegree has a similar growth compared to the base-2 logarithm of the network size \( N \). Qualitatively, the comparison with the result in (8) on the indegree distribution is consistent with the empirical evidence on Twitter [57] and YouTube [58] on the asymmetry between indegree and outdegree distributions. Moreover, the result quantitatively matches the empirical observations on the small-world property of real-world networks, according to which the distance between two randomly chosen nodes grows proportionally to the logarithm of network size [59]. Unlike the preferential attachment model [36] which is based on undirected networks, our quality-based model shows that scale-free indegree distributions can be compatible with a dynamics where agents have limited outdegree, given their limited time and attention in the social networks.
We find evidence of such a limited (small) outdegree in our Twitch data-sets. As shown in Fig. 8, the outdegree probability density function is concentrated in the range \( d \in [1, 10] \), with the 99-percentile being at \( d = 15 \) (\( d = 19 \)) in the chess (poker) data-set. The maximum outdegree are found to be respectively \( d = 151 \) and \( d = 142 \), which are three orders of magnitude smaller than the maximum indegree. At first glance, the theoretical and empirical results in Figs. 6 and 8 look qualitatively strikingly similar. Compared to the data, though, the empirical distribution exhibits a larger frequency of low-outdegree nodes. To better understand this inconsistency, we studied the stacked frequencies of the outdegree of the followers of each of the 15 most followed nodes in the two data-sets, shown in Supplementary Fig. 6. We conjecture that this is partly the effect of the recommendation systems behind the Twitch platform. For this reason, users that are not necessarily interested in a certain category, e.g., chess, might follow one of the top nodes, even without having performed a quality-based exploration.

\[
\Pr[d^\text{out} = d]\quad\text{Chess dataset}\quad\text{Poker dataset}
\]

![Figure 8. Outdegree probability density function for the Twitch data-sets concerning the users in the categories chess (blue) and poker (purple). Note that, in the plot, we neglect those followers with outdegree \( d = 1 \) of the top 4 nodes, whose audiences show a substantially larger (compared to the others) fraction of them. Such a discrepancy, possibly due to the effect of the recommendation systems, is discussed in Supplementary Note 3 and in Supplementary Fig. 6.

**Audience Overlap**

Another interesting analysis concerns the similarity between the followers’ sets of the different agents. Similarity between agents reveals the existence of a common interest, and it can be used for link prediction [60] or to improve the recommendation systems [61]. Inspired by the Jaccard index introduced for species similarities [62], and already used to measure followers’ overlap [63], we propose the following real-valued matrix \( O \) to measure the overlap between the audiences of two agents:

\[
O(i, j) := \frac{|\mathcal{F}_i^m \cap \mathcal{F}_j^m|}{|\mathcal{F}_i^m|} \in [0, 1],
\]

if \( |\mathcal{F}_i^m| > 0 \), and 0 otherwise, where \( \mathcal{F}_i^m \) denotes the set of followers of agent \( i \). In other words, this coefficient measures the number of common followers of \( i \) and \( j \), normalized by the number of followers of \( i \). Note that, when agent \( i \) is lower in the ranking list with respect to agent \( j \), i.e., \( i > j \), we typically have \( |\mathcal{F}_i^m| < |\mathcal{F}_j^m| \), and (10) corresponds to the Szymkiewicz-Simpson coefficient, also known as the overlap coefficient [64], where the denominator is replaced by \( \min \{ |\mathcal{F}_i^m|, |\mathcal{F}_j^m| \} \). Compared to it and to the Jaccard index [62], whose denominator is instead \( |\mathcal{F}_i^m \cup \mathcal{F}_j^m| \), our measure leads to a non-symmetric matrix.

In Fig. 9a we plot the numerical results on the overlap index for networks of 201 nodes, upon reaching the equilibrium. The results are averaged over 1000 simulations. According to the previous results, all the nodes should follow node 1 at equilibrium, thus any follower of a node \( i \) should also be a follower of node 1 and the overlap in the first column is simply \( O(i, 1) = 1 \). Moreover, consistent with our result in (8), we observe the Zipf’s sequence \( 1, 1/2, 1/3, \ldots \), in the first row \( O(1, j) = 1/j \). Perhaps surprising, a similar pattern appears in all the rows, upon averaging on a sufficiently large number of simulations. We emphasize that the results are independent on the number of nodes \( N \). An intuition for these observations is provided in Fig. 10.

To further validate our theoretical model, we perform the same analysis on the overlap among the 15 most followed nodes of our Twitch data-sets, reported in Fig. 9b (for the chess data-set) and in Supplementary Fig. 7 (for the poker data-set). From a qualitative comparison, our numerical results are well aligned with the real-world data with respect to the horizontal decrease of the overlap index. Yet, the Twitch data-sets show that the overlap index is not always uniform across the different rows. For instance, in the chess data-set low-ranking nodes exhibit slightly higher overlap index with respect to the numerical results, conversely high-ranking nodes (especially 1 and 3) show the opposite behavior. A possible explanation for this phenomenon can be found on the above-average outdegree of the followers of the low-ranking nodes and on the below-average outdegree of the followers of the high-ranking nodes discussed in Supplementary Note 3.

**Discussion**

The number of online platforms where users create relationships and form social networks is rapidly increasing. Most of these social networks, e.g., Instagram or YouTube, are based on User-Generated Content. Despite the interdisciplinary effort from multiple communities in modelling and analyzing the process of social network formation, scarce attention has been devoted to the different type and quality of this content as leading mechanism. To overcome this limitation, we proposed a quality-based network formation model in which actors share a common interest and aim at optimizing the quality of the received content by strategically choosing their followees. To support this meritocratic principle, we provided empirical evidence from longitudinal Twitter data.

We then statistically described the properties of the resulting networks. First, we discovered that the expected indegree as
This overlap is symptomatic of the homophily among the nodes, hence the result of a common interest.

Finally, to validate our model, we collected different data-sets from the popular gaming platform Twitch, and we provided strong evidence that our theoretical statistical properties are compatible with empirical network data. Despite being of simple and parsimonious form, our quality-based model already captured a number of macroscopic features of today’s UGC-based online social networks, e.g., scaling property of the indegree distribution, asymmetry between indegree and outdegree distributions, followers’ sets similarities.

Thanks to its simplicity, the model can be extended in different directions, e.g., by considering different update rules or enriching it with well-known sociological incentives, e.g., network closure. Another possibility is to introduce multi-dimensional quality attributes, to cope with the possibility of multiple interests. This may lead to an interesting analysis of the competitive structural advantage of nodes with diversified audience. Another direction consists in the theoretical and numerical analysis of the information spreading characteristic of the resulting networks, with particular emphasis on the network influencers [65]. Last but not least, future work could address the role of the recommendation systems acting on the social media platforms as well as their effect on the users’ behavior. The interplay between users’ behavior and platforms mechanisms represents a widely unexplored research direction that may shed light on the effect of digitalization on our societies.
**Theoretical Analysis**

The goal of our theoretical analysis is to derive the macroscopic statistical properties of the network resulting from the quality-based network formation model. Our first result concerns the convergence to an equilibrium state, as in the following theorem.

**Theorem (Convergence).** For any set of qualities \( \{q_1,\ldots,q_N \} \), the network reaches an equilibrium almost surely.

**Proof.** Let \( t > 0 \) and \( U(t) \) be a Bernoulli random variable such that \( U(t) = 1 \) if the network formation dynamics has reached an equilibrium within \( t \) time-steps, and 0 otherwise. Since potential connections are uniformly randomly selected, the probability that an agent \( i \neq 1 \) has not found agent 1 (or that agent 1 has not found agent 2) within \( t \) time-steps is equal to \( (1 - \frac{1}{N-1})^t \). Thus, since the dynamics of the individuals are independent from each other,

\[
P[U(t) = 1] = \left(1 - \left(1 - \frac{1}{N-1}\right)^t\right)^N \to 1, \quad \text{as } t \to \infty.
\]

The proofs of the results on the indegree and outdegree probability distributions are discussed in the Supplementary Note 1.

**Experimental setup**

In order to validate the statistical results of our quality-based model, we collected three data-sets on Twitch, an online social media platform focusing on video streaming that recently became extremely popular among gamers. Twitch users can create their dedicated channels to stream their game-play. Their UGC, in the form of live-streaming, can be browsed in appropriate categories corresponding to specific games. Thus, users can watch the streamed content of others, and eventually become followers.

Dealing with complex real-world networks poses several problems. In particular, systems are continuously changing not just in terms of network ties, but also with new nodes (users) joining and leaving the networks. To specifically validate our model results, we need (i) first to identify a suitable category of common interest, and (ii) second to reconstruct the social network among the users that show interest for this category. According to our modelling assumptions, the system is closed with respect to the set of users, and the network formation process is a consequence of users’ interest in to a specific topic. In the context of Twitch, this requires that the set of users interested in one (and only one) specific game, or topic, is fixed over time.

In order to minimize the chance of user’s interest lability, we restricted our crawling setup to the users streaming and watching in one of the following three categories: chess, poker, and art (see Supplementary Fig. 1 for the analysis on the historical trend). Furthermore, we filtered our data by language, retaining only the English-speaking users that constitute the vast majority (see Supplementary Fig. 3). In this way, we avoid the possibility of multiple overlapping “coherent” sets (see Supplementary Note 2). Finally, we used an “interest” index to retain only those users that consistently stream in one of the chosen categories and to filter out those that may have accumulated audience because of streaming in other categories (see Supplementary Fig. 4). Based on the results of this criterion, we decided to exclude the data-set related to the category art (see Supplementary Fig. 5 and related discussion in Supplementary Note 2).

We then set up the two Twitch data crawling on the categories chess and poker. On Twitch, not all the users provide their UGC. Moreover, it is possible to crawl data only on the users that are currently live-streaming. Thus, we repeated our crawling every hour for a period of one week (starting on September 20, 2020), until reaching a stable and consistent ranking in the list of the top 30 broadcasters. We found 305 unique users streaming in the chess category, and 358 in the poker one. We then crawled the followers of these users, which we call broadcasters, obtaining a total of 690’917 (708’443) unique followers and 1’450’403 (1’739’712) ties directed towards the broadcasters in the chess (poker) category. Finally, we reconstructed the bipartite networks of broadcasters and followers as shown in Supplementary Fig. 2. Overall, the time-span of the network formation processes underlying the data-sets is comprised between April 2010 and September 2020.

**Data Availability**

The Twitter data-set and the Twitch data-sets are respectively available at the following public repositories \([39, 44]\).

**References**

1. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, 519–528 (2012).

2. Bandiera, O. & Rasul, I. Social networks and technology adoption in northern mozambique. *The economic journal* **116**, 896–902 (2006).

3. Maher, C. A. *et al.* Are health behavior change interventions that use online social networks effective? a systematic review. *J. medical Internet research* **16**, e40 (2014).

4. Robertson, S. P., Vatrapu, R. K. & Medina, R. The social life of social networks: Facebook linkage patterns in the 2008 us presidential election. In *Proceedings of the 10th annual international conference on digital government research: Social networks: Making connections between citizens, data and government*, 6–15 (2009).

5. Hall, W., Tinati, R. & Jennings, W. From brexit to trump: Social media’s role in democracy. *Computer* **51**, 18–27 (2018).

6. Jackson, M. O. A survey of network formation models: stability and efficiency. *Group Form. Econ. Networks, Clubs, Coalitions* 11–49 (2005).

7. Kaplan, A. M. & Haenlein, M. Users of the world, unite! the challenges and opportunities of social media. *Bus. horizons* **53**, 59–68 (2010).
8. Obar, J. A. & Wildman, S. S. Social media definition and the governance challenge—an introduction to the special issue. Obar, JA Wildman, S. (2015). Soc. media definition governance challenge: An introduction to special issue. Telecomm. policy 39, 745–750 (2015).

9. https://www.dsayce.com/social-media/tweets-day/.

10. https://www.omnicoreagency.com/instagram-statistics/.

11. Natali, F. & Zhu, F. A comparison of fundamental network formation principles between offline and online friends on twitter. In International Conference and School on Network Science, 169–177 (Springer, 2016).

12. Malthouse, E. C., Calder, B. J., Kim, S. J. & Vandenbosch, M. Evidence that user-generated content that produces engagement increases purchase behaviours. J. Mark. Manag. 32, 427–444 (2016).

13. Tirunillai, S. & Tellis, G. J. Does chatter really matter? dynamics of user-generated content and stock performance. Mark. Sci. 31, 198–215 (2012).

14. Choi, B. & Lee, I. Trust in open versus closed social media: The relative influence of user-and marketer-generated content in social network services on customer trust. Telematics Informatics 34, 550–559 (2017).

15. Goh, K.-Y., Heng, C.-S. & Lin, Z. Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. Inf. Syst. Res. 24, 88–107 (2013).

16. Gillin, P. The new influencers: A marketer’s guide to the new social media (Linden Publishing, 2007).

17. Jin, S. V., Muqaddam, A. & Ryu, E. Instafamous and social media influencer marketing. Mark. Intell. & Plan. (2019).

18. www.emarketer.com.

19. https://www.weforum.org/agenda/2019/10/social-media-use-by-generation/.

20. Kiss, C. & Bichler, M. Identification of influencers—measuring influence in customer networks. Decis. Support. Syst. 46, 233–253 (2008).

21. Ryu, M.-H., Kim, S. & Lee, E. Understanding the factors affecting online elderly user’s participation in video ucc services. Comput. Hum. Behav. 25, 619–632 (2009).

22. Robertson, M. Instagram Marketing: How to Grow Your Instagram Page and Gain Millions of Followers Quickly with Step-by-step Social Media Marketing Strategies (2018).

23. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. Annu. review sociology 27, 415–444 (2001).

24. Ghosh, A. & McAfee, P. Incentivizing high-quality user-generated content. In Proceedings of the 20th international conference on World wide web, 137–146 (2011).

25. Snijders, T. A. Stochastic actor-oriented models for network change. J. mathematical sociology 21, 149–172 (1996).

26. Jackson, M. O. Social and economic networks (Princeton university press, 2010).

27. Snijders, T. A. The statistical evaluation of social network dynamics. Sociol. methodology 31, 361–395 (2001).

28. Jackson, M. O. & Wolinsky, A. A Strategic Model of Social and Economic Networks. J. Econ. Theory 71, 44–74 (1996).

29. Pagan, N. & Dörfler, F. Game theoretical inference of human behavior in social networks. Nat. communications 10, 1–12 (2019).

30. Phua, J., Jin, S. V. & Kim, J. J. Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of facebook, twitter, Instagram, and snapchat. Comput. human behavior 72, 115–122 (2017).

31. Teng, S.-Y., Yeh, M.-Y. & Chuang, K.-T. Toward understanding the mobile social properties: An analysis on Instagram photo-sharing network. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 266–269 (IEEE, 2015).

32. Erdős, P. & Rényi, A. On random graphs, i. Publ. Math. (Debrecen) 6, 290–297 (1959).

33. Caldarelli, G. Large scale structure and dynamics of complex networks: from information technology to finance and natural science, vol. 2 (World Scientific, 2007).

34. Latora, V., Nicosia, V. & Russo, G. Complex networks: principles, methods and applications (Cambridge University Press, 2017).

35. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. Rev. modern physics 81, 591 (2009).

36. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. science 286, 509–512 (1999).

37. Barabási, A.-L. & Bonabeau, E. Scale-free networks. Sci. american 288, 60–69 (2003).

38. Dewey, C. Inside the world of the ‘instafamous. www.washingtonpost.com.

39. https://git.ee.ethz.ch/pagann/2020-twitter-networkscience-dataset.

40. Zipf, G. K. Human behavior and the principle of least effort: An introduction to human ecology (Ravenio Books, 2016).

41. Zhang, Q. & Sornette, D. Empirical test of the origin of Zipf’s law in growing social networks. Phys. A: Stat. Mech. its Appl. 390, 4124–4130 (2011).

11/12
42. Cristelli, M., Batty, M. & Pietronero, L. There is more than a power law in Zipf. *Sci. reports* **2**, 812 (2012).

43. Gabaix, X. Power laws in economics and finance. *Annu. Rev. Econ.* **1**, 255–294 (2009).

44. https://git.ee.ethz.ch/pagann/twitch_dataset.

45. https://www.twitch.tv/p/press-center/.

46. Ghosh, A. Social computing and user-generated content: A game-theoretic approach. *ACM SIGecom Exch.* **11**, 16–21 (2012).

47. Axtell, R. L. Zipf distribution of US firm sizes. *science* **293**, 1818–1820 (2001).

48. Kong, J. S., Sarshar, N. & Roychowdhury, V. P. Experience versus talent shapes the structure of the web. *Proc. Natl. Acad. Sci.* **105**, 13724–13729 (2008).

49. Gabaix, X. Zipf’s law for cities: an explanation. *The Q. journal economics* **114**, 739–767 (1999).

50. W. A. Hamilton, O. G. & Kerne, A. *Streaming on twitch: fostering participatory communities of play within live mixed media* (Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014).

51. Adamic, L. A. Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Res. Center, Palo Alto, CA*.

52. Gibrart, R. Les inégalits économiques. *Sirey* (1931).

53. Simon, H. A. On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).

54. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).

55. Broideo, A. D. & Clauset, A. Scale-free networks are rare. *Nat. communications* **10**, 1–10 (2019).

56. Newman, M. E. Power laws, Pareto distributions and Zipf’s law. *Contemp. physics* **46**, 323–351 (2005).

57. Myers, S. A., Sharma, A., Gupta, P. & Lin, J. Information network or social network? the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, 493–498 (2014).

58. Canali, C., Colajanni, M. & Lancellotti, R. Characteristics and evolution of content popularity and user relations in social networks. In *The IEEE symposium on Computers and Communications*, 750–756 (IEEE, 2010).

59. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *nature* **393**, 440 (1998).

60. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Soc. networks* **25**, 211–230 (2003).

61. Rawashdeh, A. & Ralescu, A. L. Similarity measure for social networks—a brief survey. In *Maics*, 153–159 (2015).

62. Jaccard, P. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat* **37**, 241–272 (1901).

63. Easley, D. & Kleinberg, J. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, vol. 81 (2010).

64. Vijaymeena, M. & Kavitha, K. A survey on similarity measures in text mining. *Mach. Learn. Appl. An Int. J.* **3**, 19–28 (2016).

65. Duffy, B. E. Social media influencers. *The Int. Encycl. Gender, Media, Commun.* 1–4 (2020).

**Acknowledgements**

The authors gratefully acknowledge financial support from ETH Zürich.

**Author contributions statement**

Authors’ contributions: N.P., W.M. and F.D. designed research; N.P. performed research; N.P. and C.L. collected data; N.P. and C.L. performed data analysis; N.P. wrote this manuscript; W.M and F.D. edited this manuscript.
Figures

Figure 1

Results from the Twitter data-set. In purple, we plot the histogram of the probability $p_i$ as defined in (1). The median of the distribution is 0.42. In blue, we compute the same distribution upon reshuffling the temporal sequences of the connections (null hypothesis). The median of this distribution is 0.49. The two distributions are statistically significantly different (p-value of Kolmogorov-Smirnov test $< 0.001$).
Figure 2

The plot shows a network of 201 nodes upon reaching the equilibrium of the quality-based model dynamics. Nodes' color is proportional to their quality attribute \( q_i \in [0, 1] \), with darker nodes having higher quality. The size of the nodes, instead, is proportional to their indegree at equilibrium. The correlation between quality (color) and indegree (size) of the nodes is a consequence of the meritocratic principle.
Given a network of $N = 1001$ agents, the plots show the probability density functions of the indegree (derived through (5) from (4) and (7)) as a function of the rank $i$. In orange, we also plot their expected value (equations (6) and (8)). In (a), the plot refers to the theoretical results at $t = 200$, in (b), the situation upon reaching the network equilibrium. The plots are shown in log-log scale, to emphasize the Zipf’s law: $E[d_{i}^{\text{in,*}}] = N/i$. 

Figure 3
Figure 4

The log-log plot shows empirical evidences of the Zipf’s law in our Twitch data-sets. In blue, we report the data on the number of followers of the top 15 users (broadcasters) in the chess category, as well as their linear regression fit of coefficient -1.04 (Pearson coefficient: -0.98). In purple, the poker data-set shows a similar fitting coefficient of -0.98 (Pearson coefficient: -0.97). Both fits are very close to the theoretically expected Zipf’s law (of coefficient -1) shown in orange.
Figure 5

Probability density function of the nodes’ indegree in a network of $N = 201$ agents. In purple, the theoretical result is shown. In orange, we plot the numerical distribution resulting from 1000 simulations upon reaching equilibrium. In blue, we fit the numerical data with a power-law of coefficient $\alpha^* = -2.110 \pm 0.005$ using the algorithm in Clauset et al. [54]. In gray, we fit them with a lognormal distribution of parameters $m^* = 1.1329$ and $s^* = 1.0282$. 
Figure 6

Outdegree probability density function for different size $N$ of the network. Note that the support of the distributions is $[1,N-1]$ but with very quickly vanishing right tails.
Figure 7. In bold blue, we plot the expected outdegree at equilibrium as a function of the number of agents $N$ in the network. In orange, the function $\log_2(N)$ shows a similar growth. The shaded area represents the confidence interval obtained with one standard deviation computed from the outdegree distribution using the definition,

$$\text{Var} \left[ d^\text{out,*}_N \right] = \sum_{d=0}^{N-1} \left( d - \mathbb{E} \left[ d^\text{out,*}_N \right] \right)^2 \mathbb{P} \left[ d^\text{out,*}_N = d \right],$$

and the recursive formula (9).

Figure 7

Please view the figure caption in the figure.
Figure 8

Outdegree probability density function for the Twitch data-sets concerning the users in the categories chess (blue) and poker (purple). Note that, in the plot, we neglect those followers with outdegree $d = 1$ of the top 4 nodes, whose audiences show a substantially larger (compared to the others) fraction of them. Such a discrepancy, possibly due to the effect of the recommendation systems, is discussed in Supplementary Note 3 and in Supplementary Fig. 6.

Figure 9
Followers’ overlap results among the top 15 nodes. In (a) the average numerical results obtained from 1000 simulations with $N = 201$ agents, upon reaching equilibrium. In (b), the results from the Twitch data-set related to the chess category. An equivalent result for the poker data-set is reported in the Supplementary Fig. 7.

Figure 10. In the sketch, we show three different possible realizations of the followers’ sets of the nodes ranking 1 – 3, for a network of general size $N$. In all these cases, the size of the followers’ sets equals the expected values, as in (8), i.e., $|\mathcal{F}_1^\text{in}| = |\mathcal{F}_1^\text{in}|/2$ and $|\mathcal{F}_3^\text{in}| = |\mathcal{F}_1^\text{in}|/3$. As one can easily verify, for all the cases, $O(i,1)=1$ and $O(1,j)=1/j$. Moreover, from left to right we have $O(2,3)$ equals to $1/3, 2/3, 0$ respectively, and $O(3,2)$ equals to $1/2,1,0$, respectively. Since the three scenarios happen with the same probability (because each user’s dynamic is an independent process), we obtain the average values of $O(2,3) = 1/3$ and $O(3,2) = 1/2$, i.e., Zipf’s law. Of course, the possible scenarios are more than the three sketched, and the sets can possibly have sizes which differ from the expected values. However, when averaging upon a sufficiently large number of realizations the audience overlap index satisfies the Zipf’s regularity in every row, as in Fig. 9a.

Figure 10

Please view the figure caption in the figure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf