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

Nicolò Pagan et al.

Automatic Control Laboratory, ETH Zürich, Physikstrasse 3 8092 Zürich, Switzerland.

pagann@control.ee.ethz.ch

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

Nicolò Pagan et al.

Automatic Control Laboratory, ETH Zürich, Physikstrasse 3 8092 Zürich, Switzerland.

pagann@control.ee.ethz.ch

The Supplementary Information is organized as follows: in Supplementary Note 1, we present the proof of the theorems on the indegree and outdegree probability density functions related to our quality-based model. In Supplementary Note 2 we describe the Twitch data-sets collection method and in Supplementary Note 3 we complement the analysis presented in the manuscript.

Supplementary Note 1. *Analytical Proofs*

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

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

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

$$\mathsf{P}\left[d_i^{\text{in}}(t)=d\right] = \tag{2}$$
$$\sum_{k=0}^{d}\binom{i-1}{k}\bar{p}_i^k\left(1-\bar{p}_i\right)^{i-1-k}\binom{N-i}{d-k}\underline{p}_i^{d-k}\left(1-\underline{p}_i\right)^{N-i-(d-k)},$$

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

$$\mathbb{E}\left[d_i^{\text{in}}(t)\right] = \frac{N}{i} - \left(\left(\frac{N-i}{N-1}\right)^t + \frac{N-i}{i}\left(\frac{N-i-1}{N-1}\right)^t\right). \tag{3}$$

Proof. First, we consider $j < i$. The event of j linking to i at time-step t only depends on the previous choices of j. If j meets i when the set of j's followees is empty (first choice), then j will certainly follow i. This happens with probability $1/(N-1)$. Otherwise, j might follow i at her k^{th} choice if and only if q_i is higher than every other q_l in the current set of j's followees. Since there are $N-i$ nodes with $q_l < q_i$, then the probability of j following i at her k^{th} choice is equal to

$$\mathsf{P}\left[a_{ji}(k)=1\right] = \left(\frac{N-i}{N-1}\right)^{k-1}\frac{1}{N-1}.$$

Then, the probability of $j < i$ following i after t time-steps is equivalent to the sum of the probability of j following i at her first, second, ..., $k^{th} \leq t^{th}$ choice, i.e.,

$$\begin{aligned}\mathsf{P}\left[a_{ji}(t)=1\right] &= \sum_{k=1}^{t}\left(\frac{N-i}{N-1}\right)^{k-1}\frac{1}{N-1} = \\ &= \frac{\left(\frac{N-i}{N-1}\right)^t - 1}{\frac{N-i}{N-1}-1}\frac{1}{N-1} = \\ &= \frac{1}{i-1}\left(1-\left(\frac{N-i}{N-1}\right)^t\right) =: \bar{p}_i(t).\end{aligned}$$

On the other hand, if $j > i$ then there are only $N-i-1$ nodes $l \neq j$ such that $q_l < q_i$. Thus, the probability of $j > i$ following i after t time-steps reads as:

$$\begin{aligned}\mathsf{P}\left[a_{ji}(t)=1\right] &= \sum_{k=1}^{t}\left(\frac{N-i-1}{N-1}\right)^{k-1}\frac{1}{N-1} = \\ &= \frac{1}{i}\left(1-\left(\frac{N-i-1}{N-1}\right)^t\right) =: \underline{p}_i(t).\end{aligned}$$

This concludes the first part of the proof. Note that $\bar{p}_i(t) = \underline{p}_{i-1}(t)$ for all $i = 2, \ldots, N$.

For the second part of the proof, note that the probability distribution of the indegree of the node i is a Poisson binomial distribution of $N-1$ independent experiments, i.e., the sum of $N-1$ independent Bernoulli trials which are not necessarily identically distributed. The j^{th} experiment, with $j = 1, \ldots, i-1, i+1, \ldots, N$, is related to the link $a_{ji}(t)$, which is a Bernoulli random variable with probability distribution described by (1). Note that $a_{ji}(t)$ is independent from $a_{ki}(t)$, for all $k \neq j$. Note also that there are $i-1$ nodes such that $j < i$, and $N-i$ nodes such that $j > i$. Thus, the probability distribution of the indegree of node i can be described as

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

where we omitted the argument t in $\bar{p}_i$ and $\underline{p}_i$.

Moreover, for a Poisson binomial distribution of n independent experiments, each with probability p_l of success, the expectation reads as $\sum_l p_l$. In this case,

$$\begin{aligned}\mathbb{E}\left[d_i^{\text{in}}(t)\right] &= \sum_{j<i} \bar{p}_i(t) + \sum_{j>i} \underline{p}_i(t) = \\ &= (i-1)\,\bar{p}_i(t) + (N-i)\,\underline{p}_i(t) = \\ &= \frac{N}{i} - \left(\left(\frac{N-i}{N-1}\right)^t + \frac{N-i}{i}\left(\frac{N-i-1}{N-1}\right)^t\right),\end{aligned}$$

which concludes the proof. □

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

$$\mathbb{E}\left[d_N^{\text{out},\star}\right] = \sum_{k=1}^{N-1} \frac{1}{k}.$$

Proof. First, we introduce the following notation $\mathbb{E}[d_N^{out,\star}] = d_N$ for the nodes' expected outdegree in a network of N agents. Here we will denote with A_N^k the event of a general node $i \in \{1,\dots,N\}$ having outdegree k at equilibrium, in a network of N agents. By definition,

$$d_N = \sum_{k=0}^{N} k\mathsf{P}\left[A_N^k\right] = \sum_{k=0}^{N-1} k\mathsf{P}\left[A_N^k\right], \tag{4}$$

where the last step derives from $\mathsf{P}\left[A_N^N\right] = 0$. Trivially we have $d_1 = 0$, and $d_2 = 1$ because in a network with only two agents, node 2 must follow (only) node 1, and node 1 must follow (only) node 2.

In the next part, we prove the following recursion formula:

$$d_N = \frac{1}{N-1}\left(\sum_{k=1}^{N-1}(d_k+1)\right). \tag{5}$$

Using the definition of d_N (4) and the formula derived in the manuscript, i.e.,

$$\mathsf{P}\left[A_N^k\right] = \frac{1}{N-1}\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right],$$

we obtain

$$\begin{aligned}d_N = \sum_{k=0}^{N-1} k\mathsf{P}\left[A_N^k\right] &= \sum_{k=0}^{N-1} k\left(\frac{1}{N-1}\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right]\right) = \\ &= \frac{1}{N-1}\left(\sum_{k=0}^{N-1} k\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right]\right).\end{aligned} \tag{6}$$

Moreover,

$$\begin{aligned}\sum_{k=0}^{N} k\sum_{l=1}^{N}\mathsf{P}\left[A_l^{k-1}\right] &= \sum_{k=1}^{N} k\sum_{l=1}^{N}\mathsf{P}\left[A_l^{k-1}\right] = \\ &= \sum_{k=1}^{N} k\left(\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right] + \mathsf{P}\left[A_N^{k-1}\right])\right) = \\ &= \sum_{k=0}^{N} k\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right] + \sum_{k=0}^{N-1}(1+k)\mathsf{P}\left[A_N^k\right].\end{aligned}$$

Note that, for all N,

$$\sum_{k=0}^{N-1}\mathsf{P}\left[A_N^k\right] = 1, \tag{7}$$

then we can use (4) and (7) to obtain:

$$\begin{aligned}\sum_{k=0}^{N} k\sum_{l=1}^{N}\mathsf{P}\left[A_l^{k-1}\right] &= \sum_{k=0}^{N} k\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right] + d_N + 1 = \\ &= \sum_{k=0}^{N-1} k\sum_{l=1}^{N-1}\mathsf{P}\left[A_l^{k-1}\right] + d_N + 1,\end{aligned} \tag{8}$$

where, for the last step, we also used the fact that $\mathsf{P}\left[A_l^{N-1}\right] = 0$ for all $l \leq N-1$.

Finally, we can use (6) and (8) to obtain

$$d_N = \frac{1}{N-1}\left(\sum_{k=0}^{N-2} k\sum_{l=1}^{N-2}\mathsf{P}\left[A_l^{k-1}\right] + d_{N-1} + 1\right).$$

Note that, for $N = 2$ the above formula correspond to (5). Instead, for $N > 2$, one can iteratively apply (8) until reaching the initial step. Therefore we have proved the recursive formula:

$$d_N = \frac{1}{N-1}\left(\sum_{k=1}^{N-1}(d_k+1)\right).$$

Next, we show that

$$d_{N+1} = d_N + \frac{1}{N}.$$

Using the recursive formula (5) we have just proved,

$$\begin{aligned}d_{N+1} &= \frac{1}{N}\sum_{k=1}^{N}(d_k+1) = \\ &= \frac{1}{N}\left(\sum_{k=1}^{N-1}(d_k+1) + (d_N+1)\right) = \\ &= \frac{1}{N}\left(\sum_{k=1}^{N-1}(d_k+1) + \frac{1}{N-1}\sum_{k=1}^{N-1}(d_k+1) + 1\right) = \\ &= \frac{1}{N}\left(\frac{N}{N-1}\sum_{k=1}^{N-1}(d_k+1) + 1\right) = \\ &= \frac{1}{N-1}\sum_{k=1}^{N-1}(d_k+1) + \frac{1}{N} = \\ &= d_N + \frac{1}{N}.\end{aligned}$$

Finally note that for $N = 2$, $d_N = 1 = \sum_{k=1}^{N-1} \frac{1}{i}$. Assume that $d_{N'} = \sum_{k=1}^{N'-1} \frac{1}{i}$ for all $N' < N$, then by induction we prove that

$$d_N = d_{N-1} + \frac{1}{N-1} = \sum_{k=1}^{N-2} \frac{1}{i} + \frac{1}{N-1} = \sum_{k=1}^{N-1} \frac{1}{i},$$

which concludes the proof. □

Supplementary Note 2. *Twitch data collection*

Twitch is an online social media platform focusing on video streaming, including broadcasts of gameplay, e-sports competitions, and real-life content. Since its launch in 2011, Twitch has gradually become one of the most popular social media platforms, and it is particularly diffused among the young generations: more than half of the audience is aged between 18 and 34, and 14% is between 13 and 18 [1]. Similarly to YouTube, the UGC on Twitch is in the form of videos, and more precisely of live-streamed videos. Users can create their dedicated channels to stream their performances that others can watch and follow. Given that only a minority of the users provide UGC, we distinguish between *streamers*, or *broadcasters*, i.e., users that produce UGC, and *viewers*, i.e., followers that consume the UGC of the streamers.

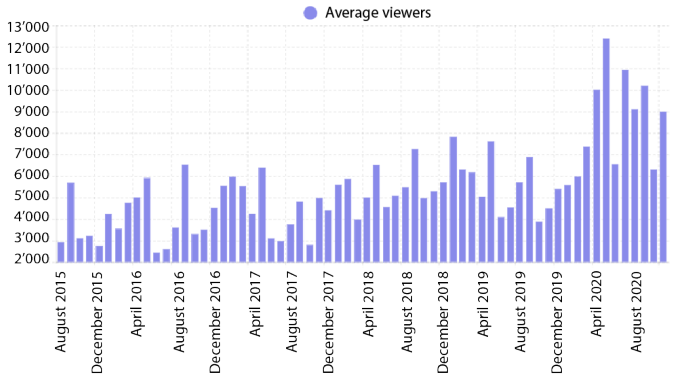
According to the topic, the content is classified into appropriate categories. Some of them are broad categories, e.g., art and music, while others are more focused, e.g., Fortnite or Minecraft (which are popular e-games). Categories help users to browse the content of their interest and to discover new streamers to follow. At the same time, categories are helpful in the identification of network communities, i.e., groups of users (streamers and their followers) with the same interest. Given the assumptions of our model, this characteristic is particularly convenient, as it simplifies the restriction of the analysis to a subset of similar users. Moreover, similar to Twitter, Twitch has the advantage that it provides its own API [2], which allows to efficiently collect data about users and their underlying interconnections. Compared to Twitter, though, Twitch API has much less restrictions in terms of rate-limits for web-crawling, thus it is more suitable for collecting large data-sets.

Category identification

In order to validate our model results, we need to collect data-sets that fulfill our assumptions. In particular, we need to find a set of users with a common interest, and to reconstruct the network of relationships among them. To do so, we first need to identify categories that define a stable community of interested users, over time.

On Twitch, more than thousand different games are streamed simultaneously. Many new games are launched every day, and users are likely to be attracted by them. Yet, most of the online games tend to have a short life-time, lasting only for a few months. Unlike them, more traditional games such as poker or chess attracted people from all over the world for centuries. Similarly, categories such as art are intuitively less subject to a change of user's interest.

To support our intuition we looked at the historical trend of viewers in the category poker, shown in Fig. 1. Until 2019, the monthly amount of viewers shows a stationary trend (except for periodic annual oscillations), indicating that the poker category is not sensitive to user's interest lability. From the comparison with several other categories, the growth during 2020 should not suggest a recent raising interest in the game of poker. Rather, it must be seen as a side effect of the quarantines due to the Covid-19 pandemic, which triggered a generalized increase in the usage of the Twitch platform [3]. Qualitatively similar results are obtained for the chess and art categories, hence they all represent suitable choices for our data-sets collection.



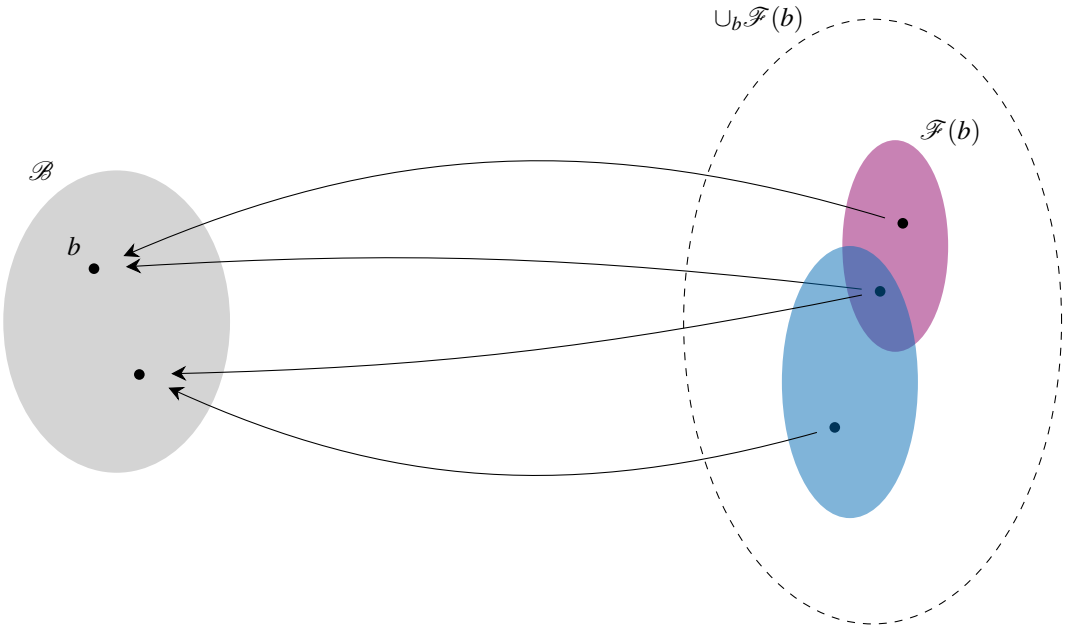
Supplementary Figure 1. Average viewers in the poker category of Twitch. Source: [4].

Crawling procedure

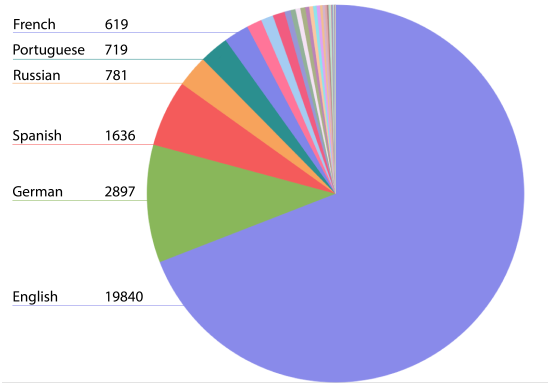
After selecting the categories, we then set up the crawling of the corresponding Twitch data-sets. For each of this category then:

- we crawl the set of all the broadcasters, or streamers, denoted with $\mathscr{B}$, that "consistently" stream into that category;
- for each user $b \in \mathscr{B}$, we crawl the set of followers of b, that we denote as $\mathscr{F}(b)$;
- finally, we construct the bipartite network where the set of nodes $\mathscr{N}$ is given by $\mathscr{N} = \mathscr{B} \cup (\cup_b \mathscr{F}(b))$, and the ties are directed from $\cup_b \mathscr{F}(b)$ to $\mathscr{B}$, as sketched in Supplementary Fig. 2.

Twitch is a multilingual platform, and broadcasters using different languages target different audiences, therefore they will unlikely form a single community. Rather, they might form a network that can be easily partitioned according to the language. Hence, we decide to restrict our crawling to the data concerning the users streaming in English, which is the most represented language, as shown for the chess category in Supplementary Fig. 3. We emphasize that, on Twitch, it is possible to crawl data only on the users that are currently live-streaming. Thus, we repeat our crawling every hour for a period of one week, until reaching a stable and consistent ranking in the list of the top 30 most followed broadcasters. At the end, we find 492, 547, and 5086



Supplementary Figure 2. Sketch of the bipartite network composed of the set $\mathscr{B}$ of broadcasters (on the left) and their followers (on the right). For each broadcaster b, we crawl the entire set of followers $\mathscr{F}(b)$. Clearly, followers' sets can overlap.



Supplementary Figure 3. Language of broadcasters which played chess between September 2019 and September 2020. Source: [5].

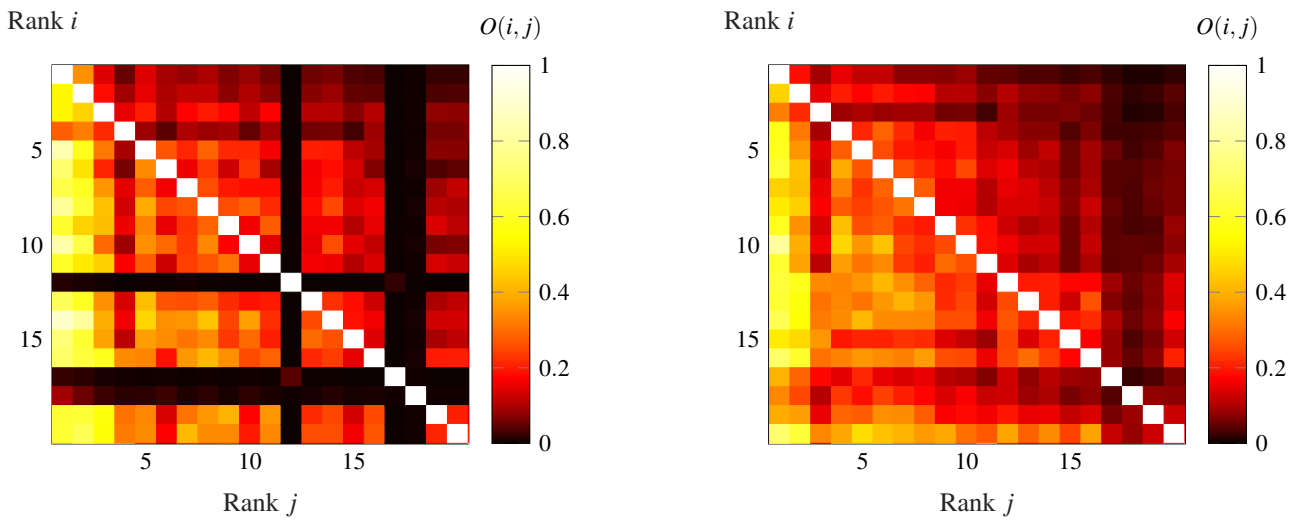
unique users streaming, respectively, in the chess, poker, and art category.

Once we collect the broadcasters' sets $\mathscr{B}$ (one for each category), we then use the Twitch API to crawl the broadcasters' followers, and we finally construct the bipartite network. We do not crawl the information on the ties between the followers, as they do not provide any UGC.

Interest index

Crawling all the (live) broadcasters in one category, e.g., chess, guarantees that they have an interest for the game of chess. However, for some of them, this may only be a secondary interest. According to our modelling assumptions, the network formation process is a based on the shared user's interest in a specific topic. One consequence of this can be found in the high overlap among the followers' sets. Reconstructing the underlying network suffers from this multi-interest effect, where the follower relationship might be due to a different interest. In that case, the followers of a certain broadcaster may present a very low overlap with the followers of the other broadcasters. An evidence of this is shown in Supplementary Fig. 4a, depicting the audience overlap index (as defined in the manuscript) among the top 20 broadcasters in the chess data-set. Clearly, some users (ranked 12, 16, and 17) have very low followers' sets intersections with the other top broadcasters.

In order to reduce the impact of the multi-interest effect, we use a filtering criterion based on the amount of time users spend on streaming on that given category. Practically, we only consider the broadcasters that "consistently" stream into the chosen category, i.e., for at least 80% of their streaming time. The intuition is that this criterion should primarily affect those broadcasters that spend a non-negligible part of their streaming time into different categories, thus accumulating a diversified audience. Note that a consistent part of the original broadcasters sets is retained: among the original 492, 547, and 5086 broadcasters in the chess, poker, and art categories, respectively 305, 358, and 2332 of them satisfy the criterion. Yet, setting such a minimum threshold leads to a generalized improvement in the followers' overlap matrix for the chess data-set, as illustrated in Supplementary Fig. 4b. Similar results are obtained for the poker data-set. On the other hand, the result concerning the art data-set, pictured in Supplementary Fig. 5, indicates that the overlap among the followers' sets remains generally very low. One possible explanation for this effect can be found in the extreme generality of this category. While both chess and poker precisely target a community of users with an interest for a specific game, the art category comprises users



Supplementary Figure 4. Followers' overlap results among the top 20 nodes in the chess data-set. In **(a)**, the broadcasters are the pure raw data. In **(b)**, we first applied the filtering criterion on the "interest" index that results in the exclusion of some broadcasters (ranked 12, 16, and 17 on the left) with very low followers' overlap.

that may be attracted by different styles. Thus, each broadcaster's followers set has very limited overlap with the others.

The different performance in terms of audience overlap index of the three data-sets highlights the importance of having a baseline community of interested users. Apparently, the network formation process related to the art category data-set is significantly different from that of the quality-based model, therefore we focus our data analysis on the other two data-sets: poker and chess.

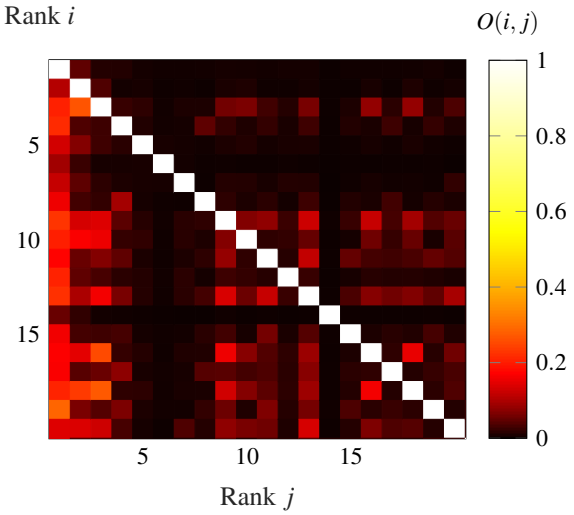
Supplementary Note 3. *Twitch data analysis*

Followers' outdegree distribution

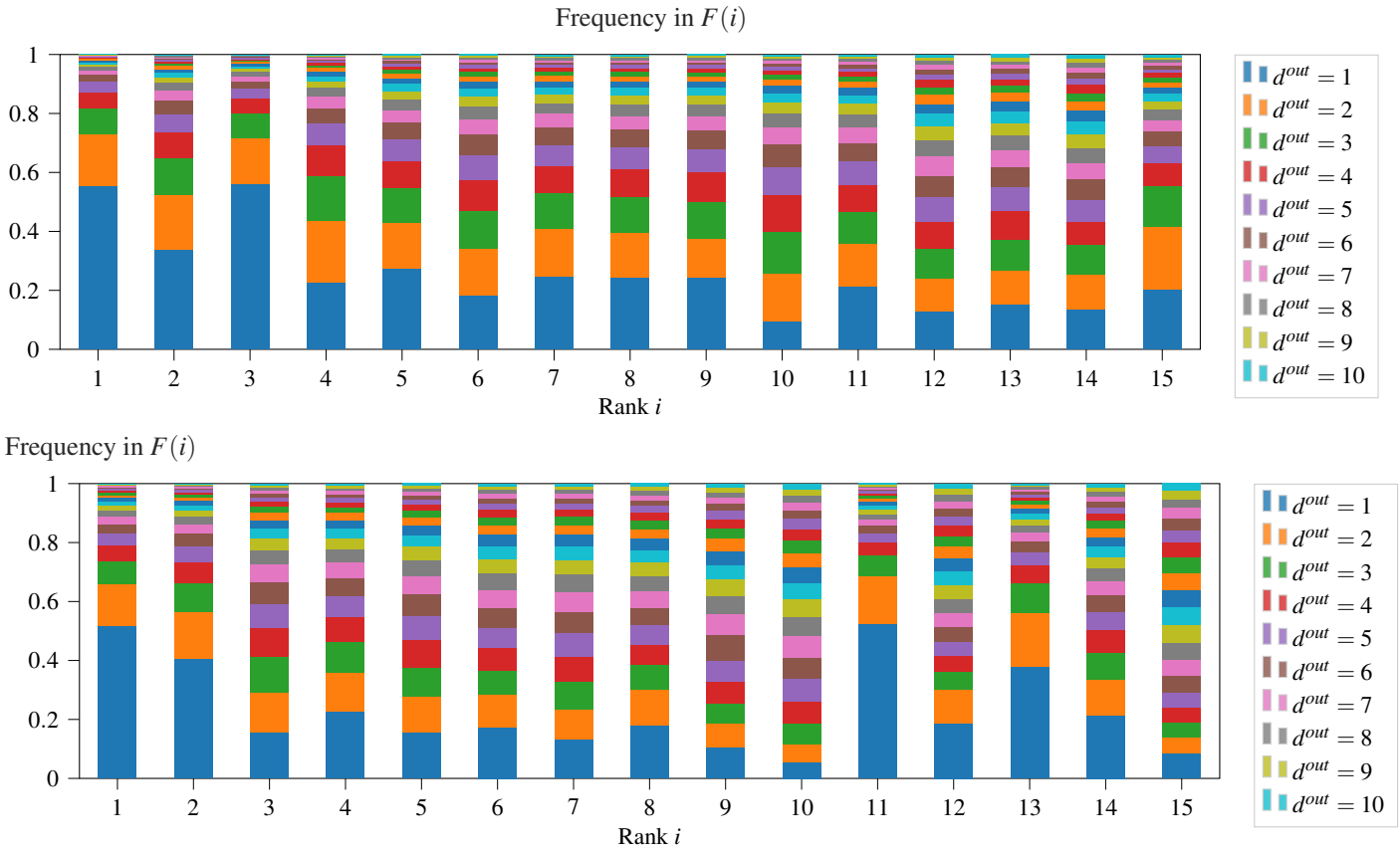
As mentioned in the manuscript, the followers' outdegree distribution found in our Twitch data-sets partly depends on the followee. Supplementary Fig. 6 shows the stacked frequencies of the outdegree of the followers of each of the 15 most followed nodes in the two data-sets. In particular, we observe that the top nodes have larger percentage of followers of outdegree $d^{\text{out}} = 1$. In other words, the top nodes are more likely to be followed by users that do not follow any other node in the category of interest. We conjecture that this is the effect of the recommendation systems behind the Twitch platform. On the other hand, a large audience with outdegree 1 can also be due to a non-perfect alignment of the node's interest, as for example for the node ranked 11 and 13 in the poker data-set (see also the audience overlap index in Supplementary Fig. 7). This can be the case of a node that accumulated followers from another category, and recently switched to playing poker. Thus, its audience is related to another category.

Audience overlap analysis

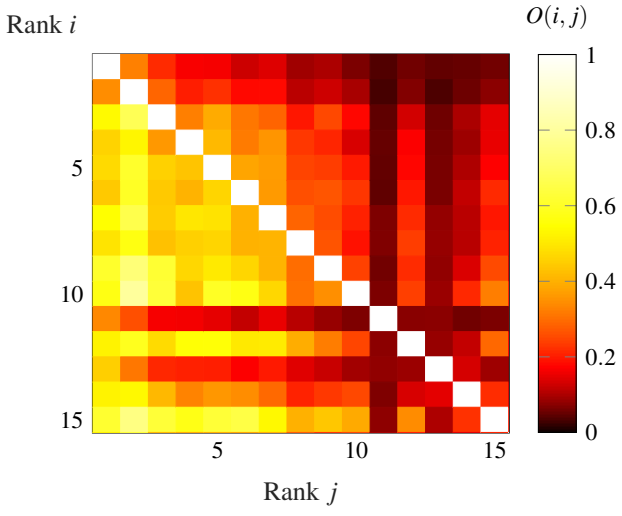
Finally, for completeness we present the results on the audience overlap index for the poker data-set (see Supplementary Fig. 7).



Supplementary Figure 5. Followers' overlap results among the top 20 nodes in the art data-set, after applying the "interest" index criterion. The result is quantitatively very different from the one presented in Supplementary Fig. 4b, denoting that the broadcasters in the art category do not share followers from the same community.



Supplementary Figure 6. The stacked histograms show, for each of the top 15 users in the two data-sets (chess on top, poker on bottom), the followers' outdegree distribution. In other words, we partition the set of followers of each of the top 15 nodes by their outdegree.



Supplementary Figure 7. Followers' overlap results among the top 15 nodes in the poker data-set.

Compared to the simulations results in Fig. 10a, here the nodes tend to share a slightly larger audience. Again, that could be the result of a recommendation system that suggests users to follow nodes similar to those they already follow.

References

1. https://twitchadvertising.tv/audience/.
2. https://dev.twitch.tv/docs/api/reference.
3. D L. King, J. B., P. H. Del Fabbro & Potenza, M. N. *Problematic online gaming and the COVID-19 pandemic* (Journal of Behavioral Addictions 9(2):184-186, 2020).
4. https://sullygnome.com/game/Poker/longtermstats.
5. https://sullygnome.com/game/Chess.