Emergence of Zipf's law among social networks influencers ⋆

Nicolò Pagan * **Wenjun Mei** ** **Florian Dörfler** ***

* *University of Zürich, Department of Computer Science, Switzerland (e-mail: nicolo.pagan@uzh.ch).*
** *Peking University, Department of Mechanics and Engineering Science, China (e-mail: mei@pku.edu.cn)*
*** *ETH Zürich, Department of Electrical Engineering and Information and Technology, Switzerland*

Abstract: User-Generated Content (UGC) is a fundamental feature of many of today's most used online social networks such as Instagram, YouTube, Twitter, or Twitch. Through the integrated search engines, users can explore the content of their peers, and those that produce higher quality UGC can attract more followers. Inspired by a meritocratic principle, we propose a novel network formation model for directed online social networks, in which actors continuously search for the best UGC provider. We theoretically and numerically analyze the properties of the resulting networks. Among other realistic network features, we found that the in-degree follows a Zipf's law with respect to the UGC quality-ranking. Furthermore, the result is robust against the effect of the recommendation systems. This extended abstract is based on Pagan et al. (2021).

Keywords: Social Networks, Dynamical systems, Social Computing, Computational Social Sciences, Data-driven decision making

1. INTRODUCTION

Since the past couple of decades, online social networks deeply affect our lives, e.g., in terms of the information we receive (Bakshy et al. (2012)), the technology we adopt (Bandiera and Rasul (2006)), the opinion we have (Hall et al. (2018)), and so on.

While there has been a coming-together of researchers from different disciplines to advance our understanding of the phenomena that take place on online social networks, these platforms continuously evolve into new forms. Compared to those that flourished in the first decade of the 21^{st} century, e.g., Facebook and LinkedIn, today's most popular platforms are directed networks that do not necessarily require friendships to be mutual. Rather than merely connecting to their real-life contacts, users of Twitter, Instagram, or TikTok prefer to use the integrated search engines to explore the content, e.g., tweets, pictures, or videos, generated by unknown peers. By doing so, they tend to become followers of real-life strangers, and to create interest-based communities that revolve around influential users that share the most interesting content.

The possibility of reaching wide audiences (way beyond real-life friends) has favoured the emergence of the so-called *new influencers* (Gillin (2007)), individuals who rapidly gain popularity by focusing on creating attractive User-Generated Content (UGC). This trend has deeply influenced consumers' and companies' behavior in markets to the point that more than 70% of US businesses engaged Instagram influencers to promote their products in 2017 [1] .

Given the potentially profound impacts of the UGC-based online social platforms on, e.g., information spreading, it is of paramount importance to understand the statistical features of these networks, especially in relation to the most influential individuals and their UGC. Here, we report some of the most relevant findings related to our recent publication, Pagan et al. (2021), in which we proposed a simple yet predictive network formation mechanism (i.e., a random graph model) based on the quality of the UGC. In our original work, we found empirical evidence from a Twitter data-set that the formation process is a result of the individuals' continuous search for better quality UGC, measured by the alignment with the follower's interests, i.e., homophily (McPherson et al. (2001)), and its goodness. Based on this sociological evidence, we assume agents are endowed with an attribute defining the quality of their UGC, and they decide their followees according to an utilitarian and meritocratic principle: they aim at optimizing the quality of the content they receive.

We analytically and numerically study the properties of the resulting networks, with a particular focus on the most influential nodes, i.e., the users with the greatest number of followers. Among other results, we observe that the in-degree distribution satisfies the well-known scale-free property (Barabási and Albert (1999)), but we also discover a specific pattern: the highest quality node

⋆ The authors gratefully acknowledge financial support from ETH Zürich and SNSF under the NCCR Automation grant, the University of Zürich, and National Natural Science Foundation of China under the grant number 72131001.

[1] www.emarketer.com

expects to have twice (respectively, three times) as many followers as the second (respectively, third) highest, and so on. This empirical regularity goes under the name of Zipf's law (Zipf (2016)), and it has been found in many real-world systems (Gabaix (2009)). Interestingly, we show that the result is robust against the effect of recommendation systems (which increase the visibility of popular nodes).

In our original work (Pagan et al. (2021)), we also show that our quality-based model predicts many interesting real-world social networks features: small diameter, small (but not vanishing) clustering coefficient, a significant overlap in the followers' sets as a result of the homophily that characterizes agents with similar interests, and a small maximum out-degree, which is consistent with the limited time users spend on these platforms. Furthermore, we validated the model predictions against three data-sets collected from Twitch, a popular platform for online gamers. This extended abstract summarizes part of our results.

2. MODEL

To formalize our quality-based model, we consider $N \geq 2$ agents whose UGC revolves around a specific common interest, e.g., a particular traveling destination. Each actor i is endowed with an attribute q_i, drawn from a probability distribution, e.g., uniform, normal, exponential distribution, that describes the average quality of i's content. As will be manifested later, our model predictions are independent of the numerical representation of these qualities, which could be somehow subjective and arbitrary. Instead, in our model, only the ordering of the individual qualities matters.

The quality q_i can be seen as the realization 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. A value of zero, instead, can be used to model users that do not produce any UGC. With this setup, the model can be directly applied to the platforms, e.g., YouTube or Twitch, in which users can be partitioned into two classes, i.e., the content creators and their followers.

Then, we consider the unweighted directed network among these agents. We denote the directed tie from i to j with $a_{ij} \in \{0, 1\}$, where $a_{ij} = 1$ means i follows j, and we assume that each agent i can only control her followees a_{ij} (excluding self-loops) but not her followers a_{ji}. We then consider a sequential dynamical process starting from the empty network, where at each time-step $t \in \{1, 2, \dots\}$ each actor i picks another distinct actor j, chosen randomly from a probability distribution on $\{1, \dots, i-1, i+1, \dots, N\}$, and decides whether to follow j or not. To reflect the meritocratic principle that emerged from our Twitter data-set, we base the tie formation decision 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.,

$$V_i(t) := \max_{j \in \mathcal{F}_i^{\text{out}}(t)} q_j, \tag{1}$$

where $\mathcal{F}_i^{\text{out}}(t) := \{j, \text{ s.t. } a_{ij}(t) = 1\}$ 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:

$$a_{ij}(t+1) = \begin{cases} 1, & \text{if } q_j > V_i(t), \\ a_{ij}(t), & \text{otherwise,} \end{cases} \tag{2}$$

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 out-degree, in Pagan et al. (2021) we show that this is not the case because in the payoff (1) the cost of good-quality connections is infinitely low, but the cost of poor-quality ones is infinitely high.

A natural question that arises when defining a dynamical process is whether it reaches or not an equilibrium. In Pagan et al. (2021) we show that an equilibrium state is reached almost surely. Without loss of generality (see also (Pagan et al. (2021))), we re-order the agents by decreasing quality, 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. If the probability distribution function that rules the meeting process is such that each agent eventually meets all the other agents almost surely, then convergence to an equilibrium state is guaranteed once every agent has found agent 1 (and agent 1 has found agent 2), see Theorem 1 in Pagan et al. (2021). Clearly, a uniform distribution for the meeting process satisfies the above assumption, because the probability that an agent does not find her best target within t time-steps is equal to $\left(\frac{N-2}{N-1}\right)^t \to 0$, as $t \to \infty$. Depending on the platform which is considered, though, one might expect that the probability of finding a node j depends linearly on the (current) in-degree of j, in a way similar to the preferential attachment process by Barabási and Albert (1999). In other words, the popular users are suggested more frequently by the recommendation systems built in the search engines. Note, though, that contrary to the preferential attachment mechanism, in our model the link formation depends on the quality of the user according to the utilitarian and meritocratic principle defined in (2).

To cope with different degrees of recommendation and to understand their impact on the final network structure, we define a parameter $\alpha \in [0, 1]$ which controls the amount of "preferential attachment" that is introduced in the meeting process. Thus, the probability that individual i meets individual $j \neq i$ at time step t reads as:

$$\mathsf{P}\left[M_{ij}\right] = \begin{cases} \dfrac{d_j^{\text{in}}(t) + 1}{\sum_{j \neq i} d_j^{\text{in}}(t) + 1}, & \text{with probability } \alpha \\ \dfrac{1}{N-1}, & \text{with probability } 1 - \alpha, \end{cases} \tag{3}$$

where $d_j^{\text{in}}(t)$ denotes the in-degree of node j at time-step t. Clearly, if $\alpha = 0$, the probability is uniform and independent on the network evolution. On the other extreme, when $\alpha = 1$, the probability depends over time, and grows proportional to the in-degree (i.e., popularity) of the target node.

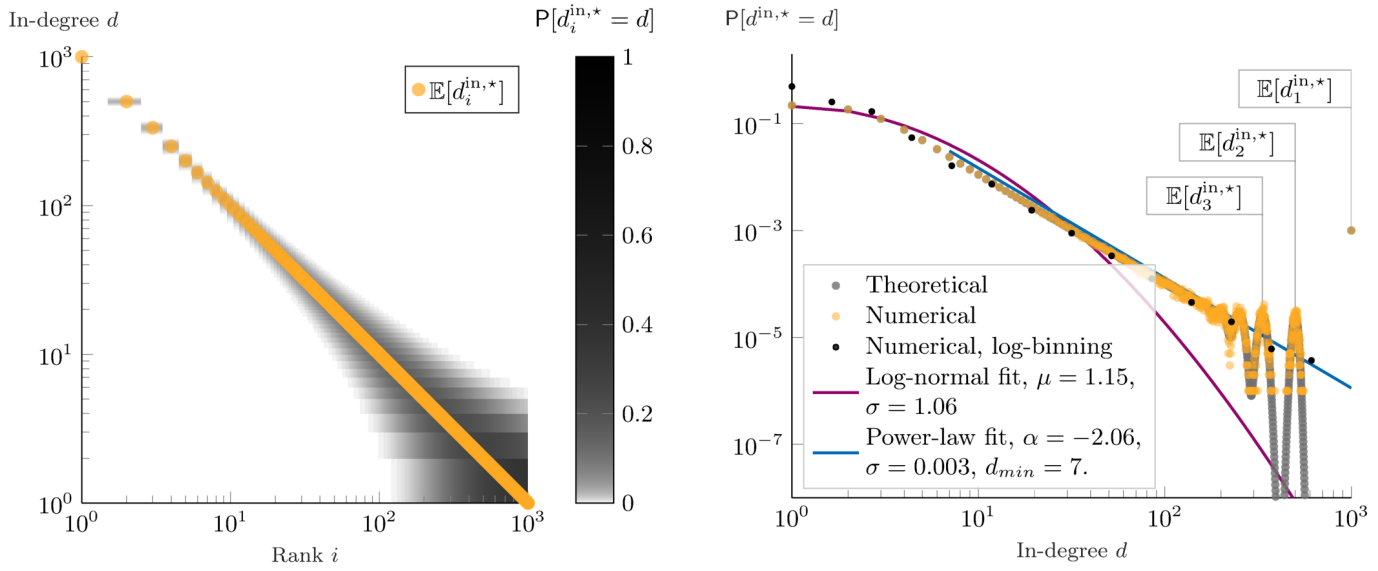

Fig. 1. Given a network of $N = 1000$ agents, the plot on the left shows the probability density functions of the in-degree with the expected value, as a function of the quality-rank. On the right, the average probability density function is plotted: in gray, the theoretical result for $N = 1000$ agents is shown; in orange, we plot the numerical distribution resulting from 1000 simulations upon reaching equilibrium. In black, the numerical distribution is shown after using the standard logarithmic binning of data. Finally, we fit the numerical data with a power-law using the algorithm in Clauset et al. (2009) (blue) and with a log-normal distribution (purple).

3. ANALYSIS

One of the most relevant aspects of social network analysis is the in-degree distribution. In fact, while other measures of centrality, e.g., closeness, betweenness, or eigenvector centrality, are also important for information diffusion, in-degree centrality immediately quantifies the reach of the content generated by a user in an online social network. For this reason, in the following theorem we study the in-degree probability density function of a network in equilibrium, under the assumption of $\alpha = 0$, i.e., uniform distribution of the meeting process. Importantly, the results are drawn as a function of the quality-rank i (for the proof, see Corollary 1 in Pagan et al. (2021)).

Theorem 1. Under the assumption of a uniform meeting process ($\alpha = 0$), at equilibrium the probability that node i is followed by node $j \neq i$ is:

$$\mathsf{P}\left[a_{ji} = 1\right] := \lim_{t\to\infty} \mathsf{P}\left[a_{ji}(t) = 1\right] = \begin{cases} \dfrac{1}{i-1}, & \text{if } j < i, \\ \dfrac{1}{i}, & \text{if } j > i, \end{cases} \tag{4}$$

and the expected in-degree of node i reads as:

$$\mathbb{E}\left[d_i^{\text{in},\star}\right] = \begin{cases} N-1, & \text{if } i = 1, \\ \dfrac{N}{i}, & \text{otherwise.} \end{cases} \tag{5}$$

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.

Such a regular scaling property is called Zipf's law (Zipf (2016)) and it is illustrated in Fig. 1 (right), where we plot the expected in-degree of each node as a function of its quality-rank, together with the probability density functions. In log-log scale, the expected in-degree perfectly matches a line with coefficient -1. Real-world evidence of Zipf's law has been reported in many systems, including firm sizes (Axtell (2001) or city sizes (Gabaix (2009)), and its peculiarity and apparent ubiquity have triggered numerous efforts to explain its origins (Gabaix (2009)). Despite being a discrete distribution, Zipf's law is often associated with the continuous Pareto distribution, better known as power-law (Adamic (2000)). However, as noticed in Cristelli et al. (2012), 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 difference with a Pareto distribution becomes evident when considering our in-degree probability density function (derived by computing the average of each user's probability density function):

$$\mathsf{P}[d^{\text{in},\star} = d] = \frac{1}{N}\sum_{i=1}^{N} \mathsf{P}[d_i^{\text{in},\star} = d]. \tag{6}$$

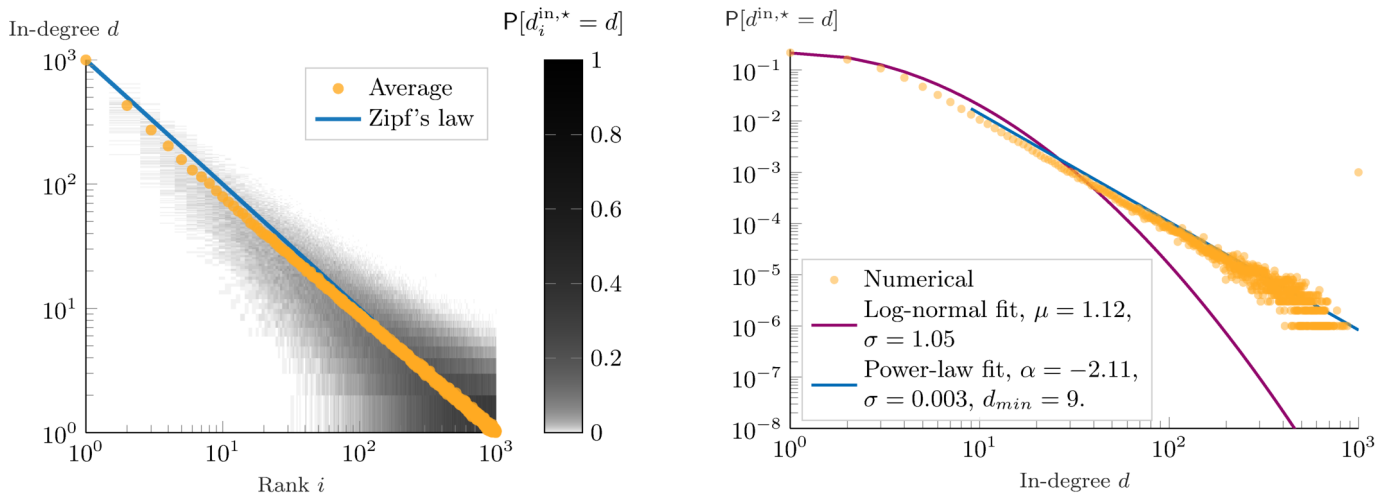

Fig. 2. Numerical results of 1000 simulations with 1000 nodes for a mixed process (with 50% probability, the potential followee is chosen from a uniform distribution, and with the remaining 50% from a preferential attachment mechanism). On the left, the in-degree probability density functions, as a function of the quality rank. On the right, the average in-degree distribution function and corresponding fits.

As shown in Fig. 1 (right), the theoretical in-degree probability density function follows a power-law of coefficient $\hat{\alpha} = -2.06 \pm 0.003$ ($p - value < 10^{-8}$), which is not surprising since the expected in-degree is distributed according to a Zipf's law (see again the discussion in Adamic (2000)). However, in our distribution we can recognize the typical Zipf's sequence, which highlights the relation between the network influencers. We refer to our original work for additional analysis and fitting comparison.

To understand the impact of the recommendation systems on the meritocratic principle, in Fig. 2 we report the numerical results obtained with $\alpha = 0.5$. Compared to the uniform distribution scenario in Fig. 1 (left), the variance of the in-degree probability distribution of each agent is increased. In this scenario, it becomes possible that some agents get an initial (i.e., in the early stage of the network formation process) advantage (or disadvantage, purely by chance), which gets reinforced by the (mixed) preferential attachment mechanism. Yet, it is remarkable that the correlation between quality and followers persists (on average): the higher the quality, the higher the average number of followers. Even more importantly, the Zipf's relation is robust under mixed preferential attachment based meeting process (e.g., recommendation systems).

4. CONCLUSION

Many of today's most popular online social networks are heavily based on User-Generated Content. Based on empirical evidence from longitudinal Twitter data, we proposed a meritocratic quality-based network formation model in which actors aim at optimizing the quality of the received content. We studied the in-degree distribution of the resulting networks, and we found that the meritocratic principle leads to a Zipf's law of the expected in-degree as a function of the quality ranking. Remarkably, the result persists notwithstanding the effect of recommendation systems.

REFERENCES

Adamic, L.A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA*.

Axtell, R.L. (2001). Zipf distribution of US firm sizes. *science*, 293(5536), 1818–1820.

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

Bandiera, O. and Rasul, I. (2006). Social networks and technology adoption in northern mozambique. *The economic journal*, 116(514), 869–902.

Barabási, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509–512.

Clauset, A., Shalizi, C.R., and Newman, M.E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661–703.

Cristelli, M., Batty, M., and Pietronero, L. (2012). There is more than a power law in Zipf. *Scientific reports*, 2, 812.

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

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

Hall, W., Tinati, R., and Jennings, W. (2018). From Brexit to Trump: Social media's role in democracy. *Computer*, 51(1), 18–27.

McPherson, M., Smith-Lovin, L., and Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1), 415–444.

Pagan, N., Mei, W., Li, C., and Dörfler, F. (2021). A meritocratic network formation model for the rise of social media influencers. *Nature communications*, 12(1), 1–14.

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