Characterizing the behavioral evolution of twitter users and the truth behind the 90-9-1 rule

Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule

Alessia Antelmi
Università degli Studi di Salerno
Fisciano, Italy
aantelmi@unisa.it

Delfina Malandrino
Università degli Studi di Salerno
Fisciano, Italy
dmalandrino@unisa.it

Vittorio Scarano
Università degli Studi di Salerno
Fisciano, Italy
vitsca@unisa.it

ABSTRACT

Online Social Networks (OSNs) represent a fertile field to collect real user data and to explore OSNs user behavior. Recently, two topics are drawing the attention of researchers: the evolution of online social roles and the question of participation inequality. In this work, we bring these two fields together to study and characterize the behavioral evolution of OSNs users according to the quantity and the typology of their social interactions. We found that online participation on the microblogging platform can be categorized into four different *activity levels*. Furthermore, we empirically verified that the 90-9-1 rule of thumb about participation inequality is not an accurate representation of reality. Findings from our analysis reveal that lurkers are less than expected: they are not 9 out of 10 as suggested by Nielsen, but 3 out of 4. This represents a significant result that can give new insights on how users relate with social media and how their use is evolving towards a more active interaction with the new generation of consumers.

CCS CONCEPTS

• **Applied computing** → **Social and behavioural sciences**; • **Human-centered computing** → *Social media.*

KEYWORDS

User behavior; Lurker; Participation Inequality; Role Discovery; Online Social Networks; Data-Driven Analysis

ACM Reference Format:
Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. 2019. Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316705

1 INTRODUCTION

The exponential growth in the use of digital devices and the ubiquitous online access allow Online Social Networks (OSNs) to represent a fertile field to collect real user data and to explore OSNs user behavior [6]. During the last decade, researchers started exploring the *evolution* of the online user activity and their OSNs social roles [10, 12]. Simultaneously, the attention toward the question of *participation inequality* began growing up, especially on the identification and profiling of passive users, aka *lurkers*. According to the 90-9-1 rule[1] about participation inequality, engagement seems not to be equally distributed among users of social media. In 2006, the expert of User Interface Jakob Nielsen introduced this rule of thumb, describing an online community as composed by 90% of lurkers, 9% by users who contribute from time to time, and only 1% by users who account for most contributions. According to this rule, all large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one property: most users do not participate and lurk in the background. Simply ignoring lurkers in social media analysis could, therefore, lead to misjudgment of overall population interests level [5]. However, due to clear privacy reasons, the identification of lurkers is not a trivial task (for example, we cannot access users' login data) and the definition of the concept of *activity* itself is not straightforward. These considerations lead to our main research question: *How can we capture and characterize OSNs user behavior and its evolution over time according to the level and typology of user interaction in online platforms?*

Our study aims to answer this question by observing the activity pattern of 122,894 Twitter users over a period of 4 months - from August 3rd to December 3rd, 2017 - and by analyzing both their visible actions (e.g., tweets, favorites) and their *hidden* activity (e.g., change of the account's screen name) on the social platform, for a final total of over 36 million actions. It is worthwhile noting that in this work the concept of activity covers all possible actions that could be done on the platform, from posting a tweet to modifying one's own Twitter account biography. The rationale behind this choice is that defining and characterizing lurkers based only on the number of tweets posted during a given time interval could be a limitation on the identification of this class of users. In this sense, our work differs from the current literature. As case study, we chose the microblogging platform Twitter because of the easiness of silent information consumption [5] and the openness of its APIs. The contribution of our work can be summarized as follows:

- Based on the data collected from the Twitter users timeline, we found that users can be identified by four levels of activity (*high*, *medium*, *low*, *no activity*).
- Our findings show that active users exhibit the same behavioral patterns despite the volume of their interactions with the social platform and that they can be described by a set of five roles - namely *Tweeter*, *Quoter*, *Retweeter*, *Replyer*, *Liker*.

The characterization of Twitter users in terms of their activity leads us towards a *corollary question* about the proportion between active and passive users within the social platform: *Does the 90-9-1 rule apply in Twitter?* While a lot of effort has been made to study the presence of this pyramid of engagement in online platforms like

[1] https://www.nngroup.com/articles/participation-inequality

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316705

blogs, forums, wikis (e.g., Wikipedia), and system reviews (e.g., Amazon) [3, 7], systematic studies have not been carried out in social networks yet.

- Our data-driven findings suggest that user engagement in social networks can be captured by our *new* four-levels classification, identifying as lurkers 3 users out of 4 - interestingly, less than the proportion suggested by Nielsen.

Some interesting works about online participation patterns and lurkers can be found in [2, 4, 5, 8, 11].

2 METHODOLOGY

Dataset. To build a dataset as heterogeneous as possible, we focused on a random subset of Twitter users. Due to the absence of Twitter APIs specifically designed to handle this task, we performed the following steps to construct our dataset.

(1) We collected a tweets sample through the Twitter streaming APIs, from which we extracted 241,000 unique users (*starting set*). We could not stop our collecting process at this point because this would have meant working on a biased dataset where all users had posted at least one status update[2].
(2) To reduce the bias introduced in the first step, we randomly selected 200 users from the *starting set*, and we collected all their followers (*followers set*).
(3) Then, we randomly picked 50 from the *followers set*. This set of 50 users will represent the *seed* of our dataset.
(4) For each user in the *seed set*, we randomly chose up to 60 his/her followers, adding them to the final dataset (*extended seed set*) and obtaining around 3,000 users.
(5) For each user in the *extended seed set* we repeated step (4), ending up with around 180,000 users.
(6) Finally, we removed all users with a private account, obtaining 131,301 unique users (*final set*).

We crawled all posts (tweets, retweets, quotes, replies), favorites, and profile snapshots of each user in the *final set* during a 4 months period, from August 3rd to December 3rd, 2017. At the end of this process, we cleaned our dataset from all users marked as spammers by the Twitter Support Team. In addition, we also removed all *churners*, typically identified as those users who are registered to an online platform/service but do not *use* it [14]. It is worthwhile to remind that lurkers *do use* to the online platform, even though they do not generate any content. To deal with this problem, we used a strategy similar to the one described by Gong et al. [5]: if a user does not perform any action during the whole period of observation, then we can consider him/her as a churner, and we definitely remove him/her from the dataset. We consider the following activities as an action: posting a new status update (tweet, retweet, quote or comment), tapping a like, following or unfollowing a new account, changing either the screen name, the description, the location, the profile image or the banner image. At the end of this process, our dataset was made up of **122,894** unique users and over **36 million** activities (posts and favorites).

[2]Twitter does not reveal how the samples are generated and does not even guarantee that the sampling ratios are stable [13].

Clustering Features. We identified two sets of features which capture the platform-specific usage and user activity. These features describe the activity-related behavior considering the time dimension (e.g., the number of activities per week) and the typology of activity done (e.g., number of tweets, likes).

To analyze Twitter users according to their *level* of activity, we represent each user with a unidimensional features vector, containing the following information extracted from the Twitter timeline of the referred user. We will refer to this set of features as **quantity-based** features, where an *activity* can be either a tweet, retweet, quote, reply and favorite.

i. **Q-Number**, total number of activities;
ii. **Q-WeeklyFrequency**, average amount of activities per week;
iii. **Q-AverageTime**, average time in days between two consecutive activities.

To study active Twitter users according to the *typology* of the visible activities done on Twitter, we represent each active user with a unidimensional features vector, containing the following information extracted from the Twitter timeline of the referred user. We will refer to this set of features as **typology-based** features.

i. **T-Tweets**, number of tweets posted;
ii. **T-Quotes**, number of quotes posted;
iii. **T-Retweets**, number of retweets posted;
iv. **T-Replies**, number of replies posted;
v. **T-Favorites**, number of favorites posted.

Workflow. *Characterizing Twitter users according to their level of activity.* Our first goal is to identify user roles based on a *quantitative* view of the user activity through a data-driven approach, without attempting to match behaviors to a pre-defined set of roles. We follow a methodology similar to the one described in the studies of O'Donovan et al. [9] and Yang et al. [14]. We employed K-means as clustering algorithm, whose scalability allowed us to analyze the entire dataset. We used the Euclidean distance as distance measure and the Silhouette analysis to automatically decide the proper number of clusters k. To confirm the significance of the cluster distributions, for each feature and for each pair of clusters we applied the ANOVA and the Tukey's HSD tests, to check whether groups in the same sample statistically differ (ANOVA) and which of them in specific have significance differences (Tukey's test). Before applying the clustering algorithm, we first standardized the feature values by removing the mean and scaling to unit variance. For this analysis, we used the **quantity-based** features. We used the opposite value for the *Q-AverageTime* variable to be semantically consistent with the other two features, *Q-Number* and *Q-WeeklyFrequency*.

Characterizing active Twitter users according to the typology of activities. The workflow just described resulted in 4 clusters, each one representing a user group with different activity levels. We focused on three of these clusters, describing users with *high*, *medium* and *low* degrees of interaction. Our purpose is to identify if active users exhibit the same behavioral patterns in terms of the typology of activity posted despite their quantity. We analyzed each cluster independently, following the same methodology illustrated in the previous paragraph. This time we used the **typology-based** set of features.

Table 1: User groups with different levels of Twitter activity.

Cluster Label	Size	Relative Size	Avg Value *Q-Number*	Avg Value *Q-Weekly Frequency*	Avg Value *Q-Average Time*
H-A	6,378	5.18%	2,282.13	133.84	1.29
M-A	13,661	11.11%	1,125.52	66.11	1.17
L-A	41,470	33.74%	100.27	5.90	2.30
No-A	61,385	49.94%	0.0	0.0	-1.0

3 RESULTS AND DISCUSSION

Twitter Users Characterization. Based on the Silhouette value, we found that 4 was the best number of clusters. Table 1 describes clusters size and average feature values.

· **High-Activity (H-A).** With a relative size of 5%, this is the smallest group found. This role is characterized by the highest average values of both *Q-Number* and *Q-WeeklyFrequency* features. This means that the users in this cluster heavily interact with Twitter, performing 19 actions on the social platform on average per day.

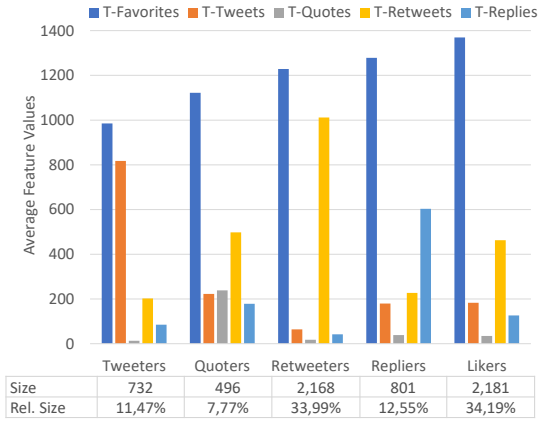
· **Medium-Activity (M-A).** This group is the most similar in terms of behavioral patterns to the *High-Activity* group, albeit with a different magnitude in terms of the overall and weekly activities done. These users interact on a daily basis with Twitter, performing 9 actions on average per day.

· **Low-Activity (L-A)** This cluster is the second larger, with a relative size of 33.74%. These users have an average weekly frequency of almost 6 actions on Twitter, and it is interesting to note that they interact with the social every two days on average.

· **No-Activity (No-A).** Unlike the previous groups, this cluster with a relative size of almost 50% is the largest and most passive group. In fact, as we can observe from Table 1, these users have done no actions over the whole four-months period of observation. The -1.0 value for the *Q-AverageTime* variable means that it was not possible to evaluate this parameter due to the absence of activities. We recall that our dataset does not contain *churners*, thus all users in this cluster have done at least one *hidden* activity.

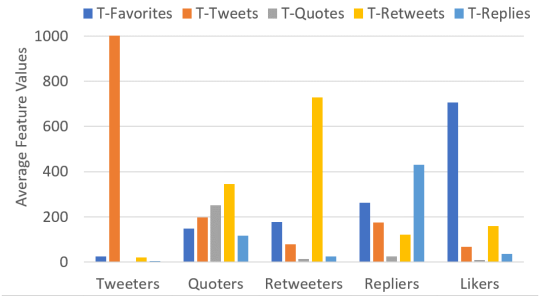
Active Twitter Users Characterization. Based on the Silhouette value, we found that 5 was the best number of clusters for the *High-Activity* and *Medium-Activity* groups; 6 was the best choice for the *Low-Activity* cluster, instead. To verify if the same set of roles occurred in each dataset, we measured the cosine similarity of every pair of role vectors. We used a threshold value equal to 0.75 to determine whether a pair matches. In practice, we found most matching pairs having a cosine similarity grater than 0.9. The sixth role identified in the *Low-Activity* dataset was not associated with any role in the other two groups. Figs. 1, 2 and 3 show the size and average feature values for each cluster. Each cluster approximately corresponds to a role and the role names were selected according to their distinguishing features, in this case, according to the type of the most *common* activity. A description of the characteristics of each cluster role follows.

· **Likers.** This role describes the behavior of the majority of the users (without considering the *Lazy* role identified in the *Low-Activity* dataset). As the label implies, the strictly predominant activity done by these users is *tapping a like*. For the *High-Activity* dataset, this typology of activity is a constant across all clusters, suggesting that the huge number of weekly actions done by these users is due precisely to this activity.



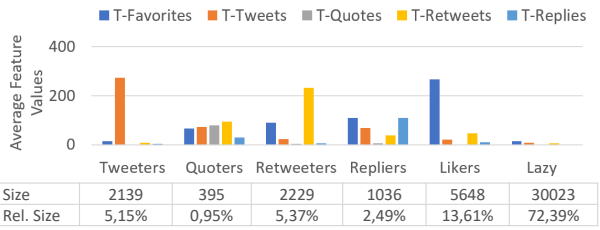
	Tweeters	Quoters	Retweeters	Repliers	Likers
Size	732	496	2,168	801	2,181
Rel. Size	11,47%	7,77%	33,99%	12,55%	34,19%

Figure 1: User roles in the *High-Activity* sub-dataset.



	Tweeters	Quoters	Retweeters	Repliers	Likers
Size	2324	477	2426	681	7753
Rel. Size	17,01%	3,49%	17,75%	4,98%	56,75%

Figure 2: User roles in the *Medium-Activity* sub-dataset.



	Tweeters	Quoters	Retweeters	Repliers	Likers	Lazy
Size	2139	395	2229	1036	5648	30023
Rel. Size	5,15%	0,95%	5,37%	2,49%	13,61%	72,39%

Figure 3: User roles in the *Low-Activity* sub-dataset.

· **Retweeters**. Second role in terms of dimension, whose main activity on the microblogging platform is *retweeting*. It is interesting to note how the second most common action in this group is liking and that, specularly, the second most common action done by the *Likers* is retweeting. Liking and retweeting do not produce any new original content; for this reason, we consider them as the most *passive* actions.

· **Repliers.** This group exhibits a behavior similar to the previous ones in terms of *liking* and *retweeting*. However, they mostly mention other users/tweets to express their opinion and share content.
· **Tweeters.** On the contrary, this group tend only to post tweets as the principal activity, without interacting with the online platform through the other typologies of action. This is especially true for the *Medium-Activity* and *Low-Activity* datasets.
· **Quoters.** In all three sub-datasets, the smallest cluster is represented by this role, whose characterizing activity is the relatively high use of *quotes*, i.e. a retweet plus a personal comment. This group is completely different from the previously described as these users seem to use Twitter homogeneously with regards to the five typologies of activities analyzed.

This second clustering analysis revealed the presence of a conspicuous group in the *Low-Activity* dataset which does not fit in the description of any of the identified roles due to the limited amount of their actions. These users - called **Lazy** - are defined by (i) a low total amount of activities (47.05 on average), (ii) less than 3 activities on average per week (2.76), (iii) *liking* as main activity and (iv) an average time between two consecutive actions of around 4 days. Despite these values, they do not share enough other distinct characteristics for the K-means algorithm to require another additional cluster to differentiate them. Analyzing the first partition obtained using the quantity-based features and a greater value for the parameter k did not directly isolate the *Lazy* group.

Does the 90-9-1 Rule Apply in Twitter? Clustering our dataset according to the *quantity-based* set of features, we ended up with four clusters. If we sort them according to their size (Table 1), we can still identify a pyramid with highly active users at the top and passive users at its bottom. This confirms the general behavior pattern theorized by Nielsen according to whom the majority of the users in an online social media are passive. The deeper analysis performed on the *Low-Activity* cluster highlights the presence of a significant subgroup (24% of the entire dataset) - the *Lazy* cluster - that significantly differs from the other five roles identified. Our interpretation is that the *Lazy* users really belong to the lurker category of *No-Active* users. In Table 2, we report the proportion of our *new* participation inequality hierarchy along with the percentages of contribution - in terms of activities - for each group. Given this consideration, our first important result is that *No-Active users are 75% of the total, thus meaning that lurkers are not 9 out 10 as suggested by Nielsen, but 3 out of 4.* This is a significant result as it can give new insights on how users relate with social media - in this specific case Twitter - and how their use is evolving towards a more active interaction with the new generation of consumers.

A comparison with Nielsen's rule. If one really wants to interpret our results to rewrite the 3-level Nielsen's rule, we can suggest some evidence to favor a 5-20-75 subdivision of the users in *Active*, *Moderately-Active* and *Silent Lurkers* groups, where: the **Active** group corresponds to the *High-Activity* cluster, the **Moderately-Active** group is the result of merging the *Medium-Activity* and the *Low-Activity* (without the subgroup *Lazy*) clusters and the **Silent-Lurkers** group coincides with the *No-Activity* cluster plus the *Lazy* subgroup (Table 3). Despite the fact that our analysis hints toward the 4-level hierarchy described in Table 2, we feel that our *AMS rule*, 5-20-75, may be the correct overall interpretation of our results.

Table 2: The new 4-layers participation inequality hierarchy.

	H-A	M-A	L-A (minus Lazy)	No-A (plus Lazy)
#Users (%)	5.18%	11.11%	9.32%	74.36%
#Actions (%)	40.14%	42.16%	12.31%	5.40%

Table 3: The *AMS* rule for Twitter.

	Active users	Moderately-Active	Silent Lurkers
#Users (%)	5.18%	20.42%	74.36%
#Actions (%)	40.13%	54.47%	5.40%

4 LIMITATIONS AND FUTURE WORK

Further investigation based on real data is needed to justify our "AMS rule", although, in our opinion, a significant step forward has been taken to ground empirical observations (like Nielsen's rule) to the real behavior of users. In this regard, we aim to validate our assumption across more Twitter datasets (e.g., a Twitter community) and domain areas, and verify their distributional results. We plan to enhance our approach - described here and in our previous work [1] - by considering not only the quantitative information but also adding the semantic of the posted content and the structural information derived from the interaction network of the users.

REFERENCES

[1] A. Antelmi, D. Malandrino, and V. Scarano. 2018. Characterizing Twitter Users: What do Samantha Cristoforetti, Barack Obama and Britney Spears Have in Common?. In *IEEE International Conference on Big Data (Big Data)*. 3622–3627.
[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2009. Characterizing User Behavior in Online Social Networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement* (Chicago, Illinois, USA) *(IMC '09)*. 49–62.
[3] J. Chan, C. Hayes, and E. Daly. 2010. Decomposing Discussion Forums and Boards Using User Roles. In *ICWSM'10*. 215–218.
[4] R. García-Gavilanes, A. Kaltenbrunner, D. Sáez-Trumper, R. Baeza-Yates, P. Aragón, and D Laniado. 2014. *Who Are My Audiences? A Study of the Evolution of Target Audiences in Microblogs*. 561–572.
[5] W. Gong, E. Lim, and F. Zhu. 2015. Characterizing Silent Users in Social Media Communities. In *ICWSM'15*. 140–149.
[6] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. 2013. Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine* 51, 9 (2013), 144–150.
[7] P. Krammer, M. Kvassay, J. Mojžiš, I. Budinská, L. Hluchý, and M. Jurkovič. 2018. Clustering analysis of online discussion participants. *Procedia Computer Science* 134 (2018), 186 – 195.
[8] M. Muller. 2012. Lurking As Personal Trait or Situational Disposition: Lurking and Contributing in Enterprise Social Media. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. 253–256.
[9] F. T. O'Donovan, C. Fournelle, S. Gaffigan, O. Brdiczka, J. Shen, J. Liu, and K. E. Moore. 2013. Characterizing user behavior and information propagation on a social multimedia network. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW'13)*. 1–6.
[10] M. Revelle, C. Domeniconi, and A. Johri. 2016. Persistent Roles in Online Social Networks. In *Machine Learning and Knowledge Discovery in Databases*. 47–62.
[11] N. Sun, P. PL. Rau, and L. Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior* 38 (2014), 110 – 117.
[12] O. Varol, E. Ferrara, C. L. Ogan, F. Menczer, and A. Flammini. 2014. Evolution of Online User Behavior During a Social Upheaval. In *Proceedings of the 2014 ACM Conference on Web Science* (Bloomington, Indiana, USA) *(WebSci '14)*. 81–90.
[13] Y. Wang, J. Callan, and B. Zheng. 2015. Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API. *ACM Trans. Web* 9, 3 (2015), 1–23.
[14] C. Yang, X. Shi, L. Jie, and J. Han. 2018. I Know You'll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application. In *KDD'18*

(London, United Kingdom). 914–922.