An Empirical Study of the Diversity of Athletes' Followers on Twitter

Ricardo Silveira, Giulio Iacobelli and Daniel Figueiredo

Abstract The study of user diversity in online social networks is an important and ongoing research effort to better understand human behavior. This work takes a step in this direction by providing an empirical study of around $8,000$ athletes divided into 13 categories and followed by 197 million users in Twitter. We propose a metric for follower diversity at the category level that factors the vast popularity difference between categories (e.g., soccer versus golf). Using this metric, we propose a measure for athlete heterogeneity based on the diversity of his/her followers. Our findings reveal that follower diversity is spread across two scales with the vast majority of users having very small diversity. We also find that athlete heterogeneity is inversely proportional to its number of followers. This indicates that very popular athletes are followed by users that (on average) do not follow other sports.

1 Introduction

Online social networks became highly popular in the last decade, and currently a massive amount of data is generated daily by the millions of users of such systems. This data has been leveraged to study various aspects of human behavior at unprecedented scale, such as information cascades, news bias, opinion formation and others [5, 6, 7]. Among the most widely used and studied platforms is Twitter, a system blending social media and social networks with over half a billion users and the advantage that in principle all user generated data is publicly available [2, 3, 4].

An interesting object of study is user interest diversity [1, 6, 8]. In systems where so much is available on so many topics will users' interest become more or less diverse? Will strong polarization around topics emerge out of local interactions? Can we accurately quantify user diversity in the presence of strong biases imposed by popularity and context? Prior works have focused primarily on user generated content to assess diversity and polarization. We focus on a much stronger signal which does not depend on content and topic categorization is external to the system. In particular, we focus on a target group where topics and interests are well-defined: sports, athletes and followers. In particular, we present an empirical study of diversity considering around 8 thousand athletes divided into 13 categories and followed by 197 million people on Twitter. Our main contributions and findings are:

Department of Computer and System Engineering (PESC), Federal University of Rio de Janeiro (UFRJ), Brazil, e-mail: [ricardosilveira,giulio,daniel]@land.ufrj.br

- Characterize athlete/follower relationship in the context of different categories.
- Propose a metric to quantify user diversity and athlete heterogeneity that accounts for popularity bias of categories.
- Quantify the proposed metrics and identify various relationships between categories, popularity and diversity.
- Among other findings, we highlight that popular categories are followed by less diverse users, and that athletes with many followers have less diverse followers.

The remainder of this paper is organized as follows. Section 2 describes the data collection and presents a preliminary analysis. Sections 3 and 4 present the proposed metrics for user diversity and athlete heterogeneity and their empirical evaluation, respectively. Section 5 concludes the paper with a brief discussion.

2 Data Collection and Analysis

The website *www.tweeting-athletes.com* is a semi-public platform keeping track of athletes that have a Twitter account. Besides associating an athlete to its twitter account, the website has categorized all athletes according to their sport or league in which they play, such as soccer and NBA. Since athlete information is manually verified by the website, the available data can be taken as reliable, although not complete since athletes (or their managers) must register with the website.

We developed a web crawler to collect all athletes' profiles on the website which on February 2015 was around $8,000$ athletes. We then developed a program to use the public Twitter API to collect information of each of these athletes, including the identity of all their followers (this procedure lasted several days, also on February 2015). The data collected is summarised in Table 1. The first column corresponds to the thirteen main categories listed in the website and each athlete is in exactly one of such categories with their respective sizes given in the third row. The fourth row denotes the number of followers (users) of all athletes in the corresponding category, while the second row shows the number of follows, which is the sum of the number of followers of each athlete in the category. Note that the number of follows is larger than the number of followers, since a follower may follow different athletes in the same category. The last column presents the *popularity* of each category, computed as the number of follows per athlete in the category.

Note that categories are of very different sizes, either when considering number of follows, number of athletes or number of followers. This reflects the enormous bias induced by popularity in such systems. To better accommodate for this bias, we consider that the popularity of each category is measured as $\#\,\text{Follows}/\#\,\text{Athletes}$, rather than simply using the number of follows or followers. In other words, the popularity of a category corresponds to the average number of follows per athlete in the category.

Figure 1 depicts the Complementary Cumulative Distribution Function (CCDF) for the degree of athletes and followers in log-log scale. For an athlete, the degree corresponds to the number of followers it has, whereas for a follower the degree

Table 1 Summary of collected data across different categories.

Category	# Follows	# Athletes	# Followers	Popularity
Soccer	649,689,003	1417	107,598,814	458,496.1
NBA	180,315,704	560	50,800,914	321,992.3
NFL	131,436,876	2187	29,500,112	60,099.2
Other Sports	114,895,505	651	37,282,225	176,490.8
Motorsports	40,263,566	122	16,808,352	330,029.2
MLB	39,629,147	658	11,505,122	60,226.7
MMA	35,901,352	334	12,056,037	107,489.1
Olympic Games	31,767,754	834	15,537,349	38,090.8
NHL	27,228,831	435	5,451,714	62,595.0
Golf	23,408,708	233	8,753,212	100,466.6
Cycling	13,536,674	151	6,407,248	89,646.8
Winter Olympics	7,582,522	222	5,125,114	34,155.5
Tennis	3,131,908	22	2,268,151	142,359.5

is the number of athletes followed. Note that 70% of all athletes have more than 8 thousand followers, with extremely popular athletes having over 10 million followers. Considering that the average number of followers per athlete is $20,552$, the degree of very popular athletes ($\approx$ 10 million) is an extremely large value, more than 400 times larger than the average value, an observation also reflected in the large standard deviation of $798,766$. The top five athletes in number of followers are listed in Table 2, along with the corresponding number of followers.

Table 2 Top five followed athletes and their number of followers. Four are soccer players.

Athlete	C. Ronaldo	Kaka	L. James (NBA)	Neymar Jr.	Ronaldinho
# Followers	33,657,773	22,092,978	18,978,751	17,008,330	11,608,006

The degree distribution (CCDF) of followers is quite different, showing that more than 80% of users follow less than 10 athletes and very few users follow more than 800 athletes (10% of the total number).[1] This much shorter tail is reflected in the average (15.7) and standard deviation (6.6) of the empirical distribution.

3 Follower's Diversity

We consider that users are interested in topics by associating topics to categories and interest by a following relationship. Thus, a user is interested in a category if he/she follows at least one athlete in that category. Note that we will not consider the

[1] Note that there is a user following all athletes - most likely an account not associated with a real person.

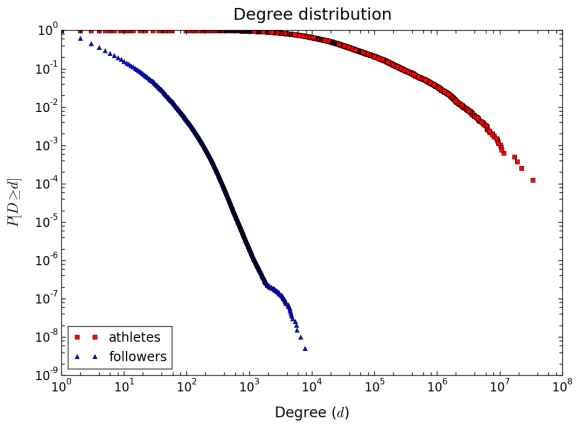


Fig. 1 Complementary Cumulative Distribution Function (CCDF) of the degree of athletes (red curve) and followers (blue curve). *Athletes' degree*: number of followers. *Followers' degree*: number of followed athletes.

number of athletes a user follows within a given category. By focusing on categories rather than on athletes we can avoid the general bias of following multiple athletes in the same category. Another important aspect when measuring diversity is the popularity of each category. As shown in Table 1, some categories are more popular than others, and to properly quantify diversity we must take this into account. Our diversity measure is build on the following observations:

- diversity should increase with the number of categories followed.
- diversity should decrease with the popularity of a followed category.

We introduce the following notation to provide a clear definition of diversity. Let the data be encoded as a directed graph $G = (V,E)$, in which a vertex $k \in V$ corresponds to a Twitter user account, and there is a directed edge $(k, j) \in E$ from k to j if user k follows user j. Let $A \subset V$ denote the set of athletes, and let S denote the set of categories. For $s \in S$, we denote by A_s the set of athletes in category s. Recall that each athlete belongs to exactly one category. A vertex $k \in V$ is a follower if there exists a $j \in A$ such that $(k, j) \in E$; we denote by $F \subseteq V$ the set of followers. Given a follower $k \in F$, and a category $s \in S$, we denote by $A_{s,k}$ the set of athletes in A_s which are followed by k, i.e., $A_{s,k} = \{j \in A_s \mid (k, j) \in E\}$. Moreover, for k a follower, we denote by $S_k = \{s \in S \mid A_{s,k} \neq \emptyset\}$, i.e., the set of categories followed by k. Given an athlete $j \in A$, we denote by $K_j = \{k \in F \mid (k, j) \in E\}$ the set of followers of j, and we denote by d_j^{in} its cardinality, corresponding to its degree, i.e., $d_j^{\text{in}} \triangleq |\{k \in F \mid (k, j) \in E\}|$. With a slight abuse of notation, given a category $s \in S$, we denote by d_s^{in} the total number of follows (of incoming links) for category s, i.e., $d_s^{\text{in}} = \sum_{j \in A_s} d_j^{\text{in}}$.

As mentioned above, to measure diversity the popularity of each category must be taken into account. Thus, we assign a weight ω_s to each category s which is

inversely proportional to its popularity, i.e., $\omega_s \triangleq \frac{|A_s|}{d_s^{\text{in}}}$. The *diversity* of a follower k, denoted by α_k, is then defined as:

$$\alpha_k \triangleq \frac{\sum_{s \in S} \omega_s \, \mathbb{I}(A_{s,k} \neq \emptyset)}{\sum_{s \in S} \omega_s} = \sum_{s \in S_k} \rho_s \,, \tag{1}$$

where, $\rho_s = \omega_s / \sum_s \omega_s$ is the normalised weight for category s, while $\mathbb{I}(\cdot)$ is the indicator function, that is $\mathbb{I}(A_{s,k} \neq \emptyset) = 1$, if $A_{s,k} \neq \emptyset$, and 0 otherwise.

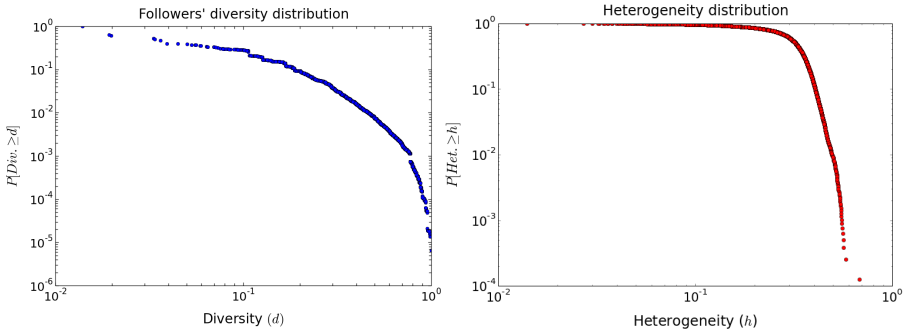
The weight and normalized weight values are listed in Table 3. Note that the diversity induced by the different categories is quite different; following Tennis contributes to user diversity much more than following NBA, while Winter Olympics contributes the most to user diversity.

Table 3 Weights (ω) and normalised weights (ρ) for each category.

Category	Soccer	Motorsports	NBA	Other Sports	Tennis	MMA	Golf	Cycling	NHL	MLB	NFL	Olympic Games	Winter Olympics
ω ($\times 10^{-6}$)	2.2	3.0	3.1	5.7	7.0	9.3	10.0	11.2	16.0	16.6	16.6	26.3	29.3
ρ ($\times 10^{-2}$)	1.4	1.9	2.0	3.6	4.5	6.0	6.4	7.1	10.2	10.6	10.6	16.8	18.7

Figure 2a depicts the CCDF of follower diversity as measured by α_k. Note that approximately 30% of the followers have the smallest possible diversity (0.014) which correspond to users who only follow Soccer category, the most popular category. Figure 2a also shows that more than 60% of the followers have diversity smaller than 0.05 and they follow athletes from one or two categories. Finally, we observe a small fraction of users (less than 10^{-5}) that follow all categories, thus having highest possible diversity of 1.

We now consider the diversity of followers of a given category with results shown in Table 4. Note that diversity is not spread uniformly across categories. In particular, followers of Winter Olympics category have the largest average diversity among all categories, with Soccer the lowest. We point out that, although following a popular category s results in a small weight ω_s, a follower may in principle follow other categories inducing a higher diversity. However, it seems that followers of popular categories do not tend to follow other categories and therefore have a small diversity. This is confirmed by considering the total and fraction of users that follow just that category (columns 4 and 5 of Table 4). Note that more than 66% of those who follow soccer, do not follow any other category, while for NFL this number is 30%.



(a) CCDF of followers diversity, α_k (avg=0.097, std=0.073).

(b) CCDF of athlete heterogeneity, β_j (avg=0.32, std=0.09).

Fig. 2 Fraction of followers/athletes with diversity/heterogeneity greater than give value (CCDF).

Table 4 Average and standard deviation of follower diversity in each category. Exclusive followers (# Excl. followers) corresponds to users who only follow that category, while their fraction is with respect to all followers of that category.

Category	Average	Std. deviation	# Excl. followers	Fraction excl. followers
Winter Olympics	0.358307	0.181506	1,428,837	0.278791
Olympic Games	0.277852	0.136397	3,499,990	0.225263
NHL	0.251917	0.175995	1,726,975	0.316777
MLB	0.248452	0.167417	3,194,130	0.277627
Golf	0.245631	0.185144	1,479,384	0.169010
Cycling	0.229478	0.181855	1,366,709	0.213307
NFL	0.200693	0.134341	8,985,095	0.304578
Tennis	0.189565	0.163223	274,485	0.121017
MMA	0.153562	0.146019	2,709,689	0.224758
Other Sports	0.126630	0.141822	15,953,684	0.427917
Motorsports	0.122397	0.144113	4,288,922	0.255166
NBA	0.114913	0.136002	18,088,226	0.356061
Soccer	0.055275	0.097337	71,148,842	0.661242

4 Athlete's Heterogeneity

We now focus on the diversity of the followers of given athletes, a concept we refer to as athlete's *heterogeneity*. In particular, we are interested in studying the relationship between the athlete heterogeneity and other characteristics, such as number of followers or popularity. Building on the concept of follower's diversity, the heterogeneity of an athlete j is defined as:

$$\beta_j \triangleq \frac{1}{d_j^{\text{in}}} \sum_{k \in K_j} \alpha_k \,, \tag{2}$$

where, K_j denotes the set of followers of athlete j and d_j^{in} its cardinality (degree). Note that β_j is the average follower diversity among the users that follow athlete j.

Figure 2b depicts the CCDF of the athlete heterogeneity. The vast majority of athletes (more than 70% of the athletes) have a heterogeneity between 0.2 and 0.4, with an average value of approximately 0.32 and standard deviation 0.09. Different from follower diversity, athlete heterogeneity is much more center around its mean with few athletes being very different. Thus, followers tend to be more diverse than an athlete's followers. Table 5 shows the top five athletes according to heterogeneity. Note that they are different from the top five according to number of followers, shown in Table 2. This already suggests that more heterogeneous athletes do not have many followers, which we next investigate. Figure 3 shows a histogram of

Table 5 Top five athletes according to heterogeneity. The most heterogeneous athlete is from Olympic Games (Volleyball) while the other four are from Winter Olympics category.

Athlete	Peter Bakare (Oly. G.)	Kim St-Pierre	Aja Evans	Molly Schaus	Kacey Bellamy
# Heterog.	0.566	0.564	0.560	0.557	0.555
# Followers	30,313	6,181	7,717	5,636	4,705

the (average) athletes' heterogeneity as a function of the logarithm of number of followers. The error bars represent the standard deviation in each bin. The figure shows a clear trend indicating that as the number of followers increases the average heterogeneity decreases. Note that, as shown in Figure 1 (red curve) there are very few athletes in the range $[0,3]$ (as in the extreme tail) and thus the data is more noisy in this range and the trend is not so clear. Overall, athlete popularity (i.e., number of followers) seems to be inversely proportional to the diversity of its followers. It is worth noting that not every athlete belonging to a very popular category has a small heterogeneity value. For example, soccer players Charlie Davies (more than 100 thousand followers) and Landon Donovan (more than 1 million followers) have heterogeneity 0.39 and 0.31, respectively. Considering that soccer weight is 0.014, this indicates that most of their followers also follow other categories.

5 Discussion and Conclusion

The study of diversity in online social networks is an ongoing research effort that can significantly contribute to the understanding of human behavior. This work takes a step in this direction by considering the context of athletes their categories and their followers in Twitter. In particular, we introduce a simple metric for follower diversity that allows a more fair comparison between categories with extremely different popularities, such as Soccer and Winter Olympics. Our analysis of around $8,000$ athletes and their 197 million followers reveals very interesting findings, such as that most followers have very small diversity (60% follow less than two categories) and that popular athletes are followed by less diverse followers. Although this work has

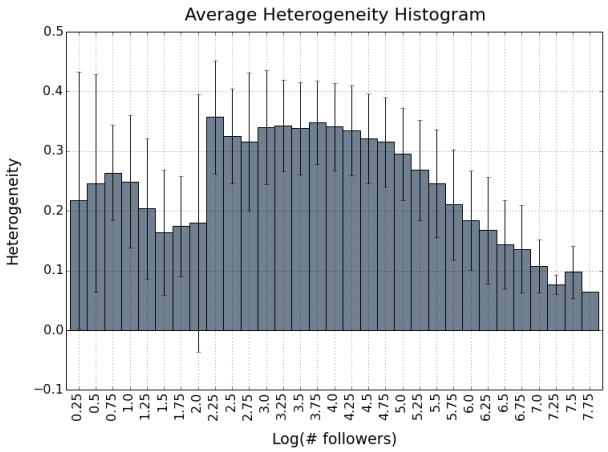


Fig. 3 Histogram of the athletes' heterogeneity as a function of log of the number of followers.

focused on diversity at the level of categories, our dataset includes subcategories for athletes, such as the team they currently play for. For future work, we intend to measure diversity considering the subcategories followed by users. Lastly, the measure hereby introduced in the context of athletes/sports can be used to study user interest diversity across different Twitter categories, such as politicians/political parties, actors/films and news/topics.

Acknowledgements This work has been partially funded through research grants from the following Brazilian agencies: CNPq, CAPES and FAPERJ.

References

1. An, J., Cha, M., Gummadi, P.K., Crowcroft, J.: Media landscape in twitter: A world of new conventions and political diversity. In: ICWSM (2011)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. ICWSM **10**(10-17), 30 (2010)
3. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: International Conference on World wide web, pp. 591–600. ACM (2010)
4. Marwick, A.E., et al.: I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. New media & society **13**(1), 114–133 (2011)
5. May, A., Chaintreau, A., Korula, N., Lattanzi, S.: Filter & follow: How social media foster content curation. In: Int. Conf Measurement and modeling computer systems, pp. 43–55 (2014)
6. Weng, L., Menczer, F.: Topicality and impact in social media: Diverse messages, focused messengers. PloS one **10**(2), e0118,410 (2015)
7. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: WWW, pp. 537–546. ACM (2011)
8. Yardi, S., Boyd, D.: Dynamic debates: An analysis of group polarization over time on twitter. Bulletin of Science, Technology & Society **30**(5), 316–327 (2010)