Characterization of Football Supporters from Twitter Conversations

Diogo F. Pacheco[†], Diego Pinheiro[†], Fernando B. de Lima-Neto[‡], Eraldo Ribeiro[§] and Ronaldo Menezes[†]
[†]BioComplex Laboratory, School of Computing, Florida Institute of Technology, Melbourne, FL, USA
[‡]Escola Politécnica, University of Pernambuco, Recife, Brazil
[§]School of Computing, Florida Institute of Technology, Melbourne, FL, USA
Email: dpacheco@biocomplexlab.org, dsilva@biocomplexlab.org, fbln@ecomp.poli.br, eribeiro@cs.fit.edu, rmenezes@cs.fit.edu

Abstract—**Football (aka Soccer) is the most popular sport in the world. The popularity of the sport leads to several stories (some perhaps anecdotal) about supporters behaviors and to the emergence of rivalries such as the famous Barcelona-Real Madrid (in Spain). Little however has been done to characterize/profile online users' behaviors as football supporters and use them as an aggregate measure to club characterization. Today, the availability of data enable us to understand at a much greater scale if rivalries exist and if there are signatures that can be used to characterize supporting behavior. In this paper we use techniques from Data Science to characterize football supporters according to their activity on Twitter and to characterize clubs according to the behavior of their supporters. We show that it is possible to: (i) rank football clubs by their popularity and fans' dislike; (ii) identify the rivalries that exist between clubs and their supporters; and (iii) find specific signatures that repeat themselves across different clubs and in different countries. The results are evaluated on a large dataset of tweets relevant to major football leagues in Brazil and in the United Kingdom.**

I. Introduction

Football (aka Soccer) is by far the most popular sport in the world with an estimate of 3.5 billion fans worldwide; Cricket, the second most popular sport, has 1 billion fewer fans. Football is played by over 250 million players and there are 209 countries recognized by FIFA (International Federation of Association Football). Such a popular sport has become part of the fabric of society and has lead to the emergence of several stories about supporters behaviors and the formation of classic rivalries. For instance, we have the Brazil–Argentina rivalry or at the club level several famous ones such as Boca Juniors–River Plate (in Argentina), Celtic–Rangers (in Scotland), Barcelona–Real Madrid (in Spain), and Palmeiras–Corinthians (in Brazil). Little however has been done to confirm such rivalries using current available data and techniques from data science. Furthermore, the characterization of online users' behaviors as football supporters may lead to a better understanding of the sport as a cultural phenomenon, be used in curtailing violence in stadiums (they are mostly due to rivalries) [1], or even used by sponsors who sometimes need to understand the behavior of supporters and their clubs [2], [3].

Data Science coupled with the availability of data can help us to unveil true rivalries, determine whether the rivalries are bidirectional, or even if the supporters' behaviors reflect the famous rivalries. In this paper, we characterize football supporters according to their activity on Twitter using a data science approach. We have identified signatures of how supporters behave and then used the signatures to classify clubs in Brazil and England according to such behaviors. Our work also brings the possibility to assess the extent to which supporter behavior is related to cultural or maybe even socio-economic factors. It becomes an open question as to whether the different behaviors are linked, are a consequence of social pressures, or is a completely independent phenomena.

This paper is divided as follows. We start in Section II with a short discussion on works related to the use of Twitter in scientific works; this section also discusses how football clubs are using Twitter to effectively engage their supporters. We then move into Section III for a description of the model we use to characterize clubs from their supporters' behavior. In Section IV we present our main findings and in particular how to cluster clubs according to the behavior of their supporters. We conclude in Section V, with some discussion of our main findings and plans for future work.

II. Related Work

Twitter is an online micro-blogging social network that has grown significantly since its foundation in 2006; as of December 2015, Twitter claimed to have 320 million active users. The user-friendly features and the powerful API for developers and scientists contribute to the growth. Twitter has become a general-content platform and it has been used to improve marketing capabilities [4], to predict political campaigns [5], to evaluate entertainment engagement [6], as a media outlet [7] and as an utility services tool [8].

In Academia, Twitter datasets are quite popular and used in social networks analyses [9], [10]. A search on Google Scholar lists over 11,000 papers published with the term ”Twitter” on the title since 2009[1]. Despite its short-length messages, Twitter data have been shown to be very rich to help researchers understand and predict human behavior. For instance, Twitter has been used to understand how collaboration emerges under catastrophic scenarios such as earthquakes [11], to track hur-

[1] Last accessed: April 25, 2016

ricanes [8], to measure the spread of happiness in a country [12], and to predict stock-market transactions [13].

Many federations, clubs, and sport players have very active profiles on Twitter[2]. For instance, the engagement of footballers changed the way clubs interact with their supporters and their players [14]. The US Soccer Federation successfully used Twitter to promote the development of the sport in the country [15]. Researchers demonstrated the capacity of real-time event detection and sentiment analysis during sporting broadcasts, such as in American football (NFL), college basketball (NCAA), and football (UEFA Champions League) [6] [16]. In addition to text, tweets can carry geo-coordinates that allows researchers to, not only understand sport events, but extrapolate and identify supporters worldwide such as in the 2014 World Cup [17].

In this work, we characterize clubs according to the behavior of their supporters in particular the supporters' attitudes towards other clubs their club play against. We show that the supporters behaviors allow us to group clubs into a few classes. Moreover, we show that we can rank clubs according to Twitter activity, and that the rank strongly correlates to established ranks of clubs. Our results suggest that Twitter helps us understand how supporters behave which may lead to more effective advertising campaigns that could be targeted to theses classes, or even a better understanding on how to ensure safety in stadiums given that rivalries have often led to violence in and around football fields. In a more targeted and up-to-the-minute manner, the real-time characterization may make standard rankings obsolete because we can gauge supporter behavior as the championships progress. It is often the case that supporters sympathize with more than one club in a championship and real-time rankings can capture the sentiment of supporters at that particular point of time as well as the dynamics of such sentiment.

III. Experiment Settings

A. Characterizing Supporters

We commence by characterizing the degree of attention given by supporters to football clubs from the frequency of mentions about a club in tweets. A *mention* in this case is the occurrence of the club's official Twitter account in a tweet. We ignore other features, such as language and location, aiming at a broader characterization. We extract these mentions about clubs from large datasets of tweets sent during competition season of major football leagues in Brazil and in the United Kingdom.

Our data are collections of tweets, where each tweet mentions one or more football clubs. Let $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ be the set of tweets sent by m users with mentions to n football clubs, where τ_i are the tweets sent by user i. We use these data to calculate a contingency table (i.e., two-way table) of frequencies of mentions users give to clubs. The contingency matrix of m users and n clubs is $U = [u_{ij}]_{m \times n}$, where an element u_{ij} is the frequency of occurrence of mentions of a club j in the tweets by user i. To avoid over counting, we treat tweets about single and multiple clubs differently. Here, we consider a tweet as a unit of user attention to a club. A club that is solely mentioned in a tweet has its user's "full attention" while multiple clubs in a tweet have the user's "divided attention". This means that if a tweet refers to C clubs, the user's attention to a club mentioned in that tweet is $1/C$. For example, in a tweet about three clubs, each club has a $1/3$ attention count while a tweet about a single club has an attention count of 1. Thus, each element in our contingency matrix is given by:

$$u_{ij} = \sum_{t \in \tau_i} \frac{\mathcal{W}_j(t)}{C_t}, \qquad (1)$$

where $\mathcal{W}_j(\cdot)$ is a function that returns the number of mentions club j receives in a tweet, and $C_t = \sum_{j=1}^{n} \mathcal{W}_j(t)$ is the number of mentions in that tweet. Finally, to ensure all users are treated equally regardless the number of tweets they send, we normalize the rows of the matrix so they sum to 1. The final normalized contingency matrix is $\hat{U} = [\hat{u}_{ij}]_{m \times n}$. For the purposes of clarity, we henceforth call $\hat{u}_{ij}$ as the *mentioned score* from i to j and it is given by:

$$\hat{u}_{ij} = \frac{u_{ij}}{\sum_{j=1}^{n} u_{ij}}. \qquad (2)$$

To characterize supporters, we further assume that a user supports one club and that the remainder clubs are *opponents*. A user's favorite club is the one to which the user mentions the most in tweets. We ignore the sentiment of tweets, so a fan can be seen as a person who follows a club most of his/her time (regardless critics), instead of someone who just expresses positive sentiment to it. We encode this user-club preference as a matrix of membership-indicator variables, $L = [l_{ij}]_{m \times n}$, where l_{ij} equals 1 for column j of the club receiving the most attention from user i, i.e., $j = \arg_j \max \hat{U}(i,j)$, and equals to 0 for all other clubs. Once the user's preferred club is known, the other clubs mentioned in tweets by that user are considered to be *opponent* clubs.

B. Characterizing Clubs

Given the above characterization of supporters, we propose indexes for characterizing club popularity and club opposition. These two indexes form the basis of our club-ranking strategy from the attention clubs receive in tweets. We define the *popularity index* of club j as the sum of all attention given to that club by either supporters or opponents, i.e.:

$$p_j = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_{ij}. \qquad (3)$$

In contrast, the club's *opposition index* is all the attention given to a club by its non supporters, which is given by:

$$o_j = \frac{\sum_{i=1}^{m} \hat{u}_{ij}(1 - l_{ij})}{\sum_{i,j=1}^{m,n} \hat{u}_{ij}(1 - l_{ij})}. \qquad (4)$$

Finally, we define a $n \times n$ club-characterization matrix $K = [k_{jk}]_{n \times n}$, where the j^{th} row is the mean of the attention given

[2] http://www.socialbakers.com/statistics/twitter/profiles/sport/

TABLE I
TWITTER OFFICIAL ACCOUNTS FROM CLUBS USED IN THIS WORK.

Brazilian “Série A” (BSA)		English Premier League (EPL)	
Club	Account	Club	Account
A. Mineiro	@atletico	Arsenal	@arsenal
A. Paranaense	@atleticopr	Aston Villa	@AVFCOfficial
Avai	@avaifc	Burnley	@BurnleyOfficial
Chapecoense	@ChapecoenseReal	Chelsea	@ChelseaFC
Corinthians	@Corinthians	Crystal Palace	@CPFC
Coritiba	@coritiba	Everton	@Everton
Cruzeiro	@Cruzeiro	Hull	@HullCity
Figueirense	@FigueirenseFC	Leicester	@LCFC
Flamengo	@Flamengo	Liverpool	@LFC
Fluminense	@FluminenseFC	Man. City	@MCFC
Goias	@goiasec_oficial	Man. Utd	@ManUtd
Gremio	@gremiooficial	Newcastle	@NUFC
Internacional	@SCInternacional	QPR	@QPRFC
Joinville	@jec_online	Southampton	@SouthamptonFC
Palmeiras	@SEPalmeiras	Spurs	@SpursOfficial
Ponte Preta	@aapp_oficial	Stoke	@stokecity
Santos	@SantosFC	Sunderland	@SunderlandAFC
Sao Paulo	@SaoPauloFC	Swansea	@SwansOfficial
Sport	@sportrecife	West Brom	@WBAFCofficial
Vasco	@crvascodagama	West Ham	@whufc_official

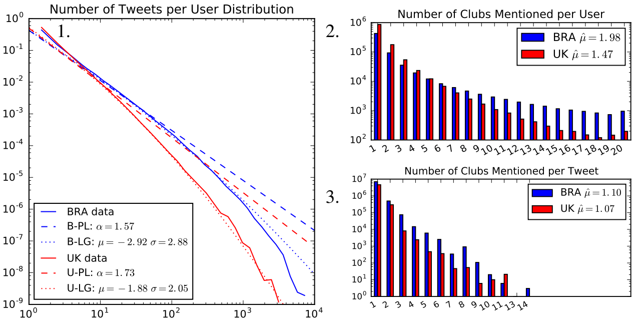

Fig. 1. Three plots describing the two Twitter datasets: BSA (in blue) and EPL (in red). 1. Distributions of number of tweets per user: real data in solid lines, log-normal in dotted lines, and power-law in dashed lines; 2. Number of clubs mentioned per user, ranging from 1 to 20 (max per league); 3. Number of clubs mentioned per tweet.

TABLE II
SOME STATISTICS ABOUT THE TWO DATASETS USED IN THIS WORK.

Statistic	BSA	EPL
Start Data Collection	25/May/15	07/Feb/15
Finish Data Collection	08/Dec/15	07/May/15
Number of Days	198	89
Number of Matches	35 / 38	12 / 38
Season Coverage Interval	8%–100%	60%–92%
Tweets with Mentions	7,578,456	4,920,764
Users with Mentions	626,208	1,151,702
Percentage of Balanced Users	6.13%	7.71%
Avg. Tweets / User	12.10	4.27
Avg. Tweets / Day	38,469	55,289
Avg. Tweets / User / Day	0.061	0.048
Avg. Mentions to Clubs / Tweet	1.10	1.07
Avg. Clubs Mentioned / User	1.98	1.47

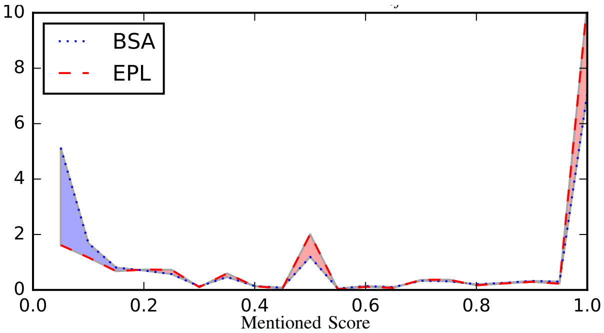

Fig. 2. Distribution of non-zero *mentioned scores* ($\hat{u}_{ij}$) for all users (Eq. 2), where the filled area highlight the greater curve.

by all supporters of club j to all other clubs. Rows of matrix K are given by:

$$\mathbf{k}_j = \frac{\sum_{i=1}^{m} l_{ij}\hat{\mathbf{u}}_i}{\sum_{i=1}^{m} l_{ij}}, \qquad (5)$$

where $\hat{\mathbf{u}}_i$ is a row in matrix $\hat{U}$. The club-characterization matrix can be written in a more compact form as $K = (L^{\mathsf{T}}L)^{-1}L^{\mathsf{T}}\hat{U}$, which simplifies its calculation.

C. Datasets

In this work we used two football datasets collected from Twitter: one representing the 2014/2015 English/Welsh Premier League[3] (EPL) and other regarding the 2015 Brazilian “Série A” (BSA). The Twitter official accounts used to identify mentions to clubs in these competitions are listed on Table I.

On Table II, we present some statistics about these two datasets. If we assume tweets and users normally distributed with time, we notice a more than 40% greater volume of tweets per day mentioning clubs in EPL than mentioning clubs in BSA. The number of users per day is 4 times greater in the former dataset. Although engagement, in terms of number of supporters, is greater in EPL, supporters of Brazilian clubs tweet 27% more per day than supporters of English clubs. In addition, on average, a single tweet contain 3% more mentions to BSA clubs than to EPL clubs (see Fig. 1.3). Overall, users on BSA dataset mention more clubs than those on EPL, 1.98 and 1.47 respectively (see Fig. 1.2). This behavior reflects on the normalized contingency matrix distribution, where BSA supporters present higher density in the lower-score region (see Fig. 2). Finally, we analyzed the distribution of the number of tweets per user in both leagues in Figure 1.1. We found that both datasets fit better on log-normal distributions than power-law and exponential distributions.

[3] The Swansea City Association Football Club, a Welsh club, plays in the Premier League. However, for simplicity reasons we call the “English” Premier League to avoid calling “English/Welsh” throughout the paper. We use the same approach when discussing clubs; we will refer to “English” clubs/supporters instead of “English/Welsh”.

TABLE III
BSA – PROPORTIONS, NUMBER OF USERS, AND RANK POSITIONS FOR CLUBS BASED ON (LEFT) POPULARITY p_j AND ON (RIGHT) OPPOSITION o_j, COMPARED AGAINST DATAFOLHA (DF14), IBOPE (I14), AND PARANÁ (P13 & P16) GREAT SUPPORTERS AND MOST HATED CLUB.

	Correlation with p_j rank			.72‡	.71‡	.83‡	.66$^\diamond$	.60†		Correlation with o_j rank			.67$^\diamond$
Club(Position)	p_j*	#Users	Rank	DF14	I14	P13	P16	Pos$^\alpha$	Club	o_j*	#Users	Rank	P16
Flamengo	17.32	108450	1	1	1	1	1	12	Corinthians	13.51	10894	1	1
Corinthians	16.73	104746	2	2	2	2	2	1	Flamengo	10.69	8619	2	2
Sao Paulo	10.42	65237	3	3	3	3	3	4	Sao Paulo	7.62	6147	3	5
Internacional	8.40	52593	4	7	9	9	10	5	Fluminense	7.39	5958	4	-
Cruzeiro	6.27	39287	5	7	7	6	6	8	Internacional	7.38	5953	5	8
Fluminense	5.93	37132	6	10	11	14	13	13	Vasco	7.23	5829	6	3
Palmeiras	5.79	36237	7	4	4	4	4	9	A. Mineiro	5.74	4632	7	6
Gremio	5.71	35760	8	6	8	7	7	3	Chapecoense	5.66	4561	8	-
Vasco	5.02	31409	9	5	5	5	5	18	Cruzeiro	5.62	4529	9	7
Santos	4.81	30122	10	7	10	8	8	7	Palmeiras	5.53	4462	10	4
A. Mineiro	4.22	26415	11	10	6	10	9	2	Gremio	5.48	4420	11	9
Sport	2.43	15224	12	12	15	11	14	6	Santos	5.08	4096	12	-
Chapecoense	2.20	13771	13	-	-	-	-	14	Sport	2.62	2110	13	-
Coritiba	1.15	7207	14	-	-	19	-	15	Figueirense	2.12	1712	14	-
A. Paranaense	1.06	6616	15	-	15	17	-	10	Coritiba	2.11	1698	15	-
Figueirense	0.89	5575	16	-	-	-	-	16	A. Paranaense	1.67	1344	16	-
Avai	0.79	4938	17	-	-	-	-	17	Avai	1.67	1342	16	-
Joinville	0.34	2123	18	-	-	-	-	20	Joinville	1.10	884	18	-
Goias	0.29	1830	19	-	-	20	-	19	Goias	0.97	783	19	-
Ponte Preta	0.24	1526	20	-	-	-	-	11	Ponte Preta	0.83	670	20	-

* Popularity and opposition indexes results in % with conf. intervals $p_j \pm 0.25\%$ and $o_j \pm 0.62\%$ at 99% conf. level. Signif. codes for ρ values: 0 ‡ .001 † .01 ◇ .05
$^\alpha$ Rank position by the end of 2015 BSA season.

IV. RESULTS

A. Mentioned Ranks

We assumed that a user supports a club if his highest mentioned score is for that club (see Section III-A). In order to check the plausibility of this assumption we ranked clubs according to our proposed indexes of popularity (Eq. 3) and opposition (Eq. 4). Then, we compared our ranks against ranks of supporters, most-hated clubs, and current season ranks using Spearman correlation [18]. The later showed the worst results suggesting that p_j captures supporters' long-term preferences.

1) Brazil: For the BSA dataset, we found four rankings regarding the size of supporters and one about most-hated clubs. They are listed bellow:

- Paraná (P16) – 4,066 people, from March to April, 2016, participated in this poll requested by GloboEsporte.com to Paraná research institute[4]. The margin of error was 1.5% at 95% confidence level. This is the only Brazilian poll to consider the most-hated clubs we had access to;
- Datafolha (DF14) – 4,337 people, in June of 2014, were interviewed by this poll from Datafolha institute[5]. Error margin was 2% at 95% confidence level;
- Ibope (I14) – 7,005 people, in 2014, participated in the poll requested by website Lance! to IBOPE research institute[6]. Error margin was 1% at 95% confidence level;
- Paraná (P13) – 7,302 people, from July to December of 2013, participated in this poll from Paraná research institute[7]. Error margin was 1% at 95% confidence level.

Since none of these ranks included all 20 clubs playing the 2015 BSA some clubs are not ranked. Table III presents the popularity and opposition shares, the prestigious ranks, and the Spearman correlation values. Although ties in the ranks are present, the difference between the spearman correlation with and without tie correction is insignificant.

The popularity index (p_j) showed high correlation values with traditional indicators of supporters size, from $r = .66$ to $r = .83$ with higher significance $p < .001$ and $p < .01$. Although the opposition index (o_j) did not correlated as high as the popularity one, o_j can still be considered a good proxy for most hated club, since $r = .67$ with enough significance $p < .05$. Furthermore, the proposed ranks present narrower confidence intervals at a higher significance level.

2) England and Wales: The comparative ranks for the English clubs were created by SportsMail on 2015[8]. Instead of using polls, they defined objective questions in order to build 7 ranks based on "hard evidence" rather than opinions: crowds, global fanbase, number of trophies, average league finish, player quality, income, and total (an average of the 6 previous ranks). We chose the 2 ranks more related to popularity, i.e. crowds and global fanbase, and the total rank to compare with our popularity index. The text of the items below were extract from their website:

- Crowds – Aggregates the rank based on the contemporary average gates during the season and the rank based on the highest biggest historic gates.
- Global Fanbase – Aggregates the total number of fans and followers from the official accounts of each club on Facebook and Twitter.

[4] http://goo.gl/4Kyqsh
[5] http://goo.gl/V5bXOz
[6] http://goo.gl/tkqEdo
[7] http://goo.gl/jRXshh
[8] http://goo.gl/xKbSu6

TABLE IV
EPL – PROPORTIONS, NUMBER OF USERS, AND RANK POSITIONS FOR CLUBS BASED ON (LEFT) POPULARITY p_j AND ON (RIGHT) OPPOSITION o_j, COMPARED AGAINST OBJECTIVE RANKS FROM SPORTSMAIL AS GREAT SUPPORTERS AND MIRROR15 AS MOST HATED CLUB.

	Correlation with p_j rank			.72‡	.94‡	.88‡	.64†		Correlation with o_j rank			.71†
Club	p_j*	#Users	Rank	Crowds	Fanbase	Total	Pos$^\alpha$	Club	o_j*	#Users	Rank	M15
Man Utd	28.39	326931	1	1	1	1	4	Man Utd	19.15	27150	1	1
Arsenal	17.78	204729	2	4	3	2	3	Liverpool	17.23	24419	2	3
Chelsea	17.81	205076	3	3	2	4	1	Chelsea	13.31	18870	3	2
Liverpool	14.10	162348	4	10	4	3	6	Arsenal	13.24	18770	4	6
Man City	6.09	70099	5	2	5	5	2	Man City	8.99	12744	5	4
Spurs	2.91	33481	6	8	6	6	5	Aston Villa	3.57	5066	6	10
Aston Villa	2.01	23170	7	9	7	8	17	Spurs	3.56	5041	7	5
Everton	1.80	20768	8	6	8	6	11	Everton	3.22	4570	8	13
Newcastle	1.45	16694	9	7	9	9	15	Newcastle	2.83	4006	9	8
West Ham	1.24	14312	10	22	10	12	12	Swansea	2.57	3644	10	18
Sunderland	0.95	10931	11	5	13	10	16	Sunderland	1.59	2249	11	11
Southampton	0.74	8487	12	34	11	15	7	Crystal Palace	1.54	2189	12	12
QPR	0.71	8227	13	36	16	31	20	West Ham	1.46	2066	13	9
Swansea	0.69	7930	14	39	12	34	8	QPR	1.42	2017	14	-
Crystal Palace	0.71	8181	15	18	19	30	10	West Brom	1.18	1667	15	13
Leicester	0.64	7320	16	21	22	21	14	Southampton	1.11	1576	16	19
West Brom	0.63	7296	17	11	20	11	13	Leicester	1.06	1508	17	13
Burnley	0.55	6386	18	27	31	26	19	Hull	0.99	1400	18	-
Hull	0.39	4543	19	19	14	32	18	Burnley	0.99	1398	18	-
Stoke	0.42	4783	20	17	17	16	9	Stoke	0.98	1395	20	6

* Popularity and opposition indexes results in % with conf. intervals $p_j \pm 0.22\%$ and $o_j \pm 0.54\%$ at 99% conf. level. Signif. codes for ρ values: 0 ‡ .001 † .01 ◇ .05
$^\alpha$ Rank position by the end of 14/15 EPL season.

In addition, we used the survey performed by the Mirror newspaper[9] on 2015 revealing the most hated clubs in EPL. We could not find further details about the number of participants or confidence margins in this poll. Table IV shows the comparative results and correlations.

Again, the popularity index seemed to have higher correlation with the "objective" ranks, ranging from $r = .72$ to $r = .94$, always with higher significance $p < .001$. Not surprisingly, the highest correlation with our method is the one based on global fanbase counts from social media. The opposition index also presented high correlation with most hated clubs in EPL ($r = .71$, $p < .01$).

B. Clubs as Distributions

Since we already presented the applicability of p_j and o_j, let us now explore club characterizations $\mathbf{k}_j$. Indeed, $\mathbf{k}_j$ allow us to understand a club j through its supporters. Moreover, it allows us to identify levels of rivalry and the main rivals themselves. In order to get a visual representation of clubs, they can be seen as histograms where each component represent a bin. There are several ways to map k_{jk} to bins, in particular, we propose:

$$\mathbf{k}_j^r : (\forall k_{jx} \forall k_{jy} \in \mathbf{k}_j)\ k_{jx} > k_{jy} \rightarrow x < y, \qquad (6)$$

where, $\mathbf{k}_j^r$ is the ranked version of $\mathbf{k}_j$ with bins sorted by mentions, i.e., the first bin represents the amount of exclusive attention users give to their own club, the second bin represents the attention given to their first rival, and so on.

Figures 3 and 5 show BSA and EPL clubs distribution, highlighting the top three rivals of each club. This straightforward representation can capture famous derbies such as: Atl. Mineiro *vs.* Cruzeiro, Flamengo *vs.* Vasco, and Grêmio *vs.* Internacional, in Brazil; Newcastle *vs.* Sunderland, Man. United *vs.* Man. City, and West Brom *vs.* Aston Villa, in England. We also noticed rivalries are not always reciprocal. For instance, Flamengo is the main rival of Fluminense, but Vasco is the rival of Flamengo; yet, Liverpool is the rival of Everton, but Man. United is the rival of Liverpool. Further, even when reciprocity exits, they can vary in intensity between sides. Apparently, Vasco and Sunderland fans oppose more their top rivals Flamengo and Newcastle than the opposite.

Moreover, clubs distributions show patterns such as height, length, and curvature which can be seen as shape signatures and in the future could be used to do a even more accurate classification of supporters for many clubs around the world. Recall that we argued in the beginning of the paper that this work could lead to better safety in stadiums; the support signature for clubs may be an indication of the level of safety for games involving teams. Moreover, these distributions are dynamic and may change from week to week. Hence, some transient rivalry could exist due to exogenous factors leading to an unsafe condition for supporters in an upcoming match. We intend to explore such issues in the near future.

C. Clustering Clubs

Grouping clubs by their ranked signature can reveal similarities on *how* their fans support instead of to *whom* they support. To measure shape similarities, distance metrics such as the Euclidean distance might be unsuitable. A better way to measure histogram similarity is to use divergence metrics, such as the Bhattacharyya distance [19].

Figures 4 and 6 show the heatmaps of a similarity matrix based on the Bhattacharyya distance. The dendrograms show the hierarchical clusters and a possible interpretation for them

[9] http://goo.gl/wVQlPO

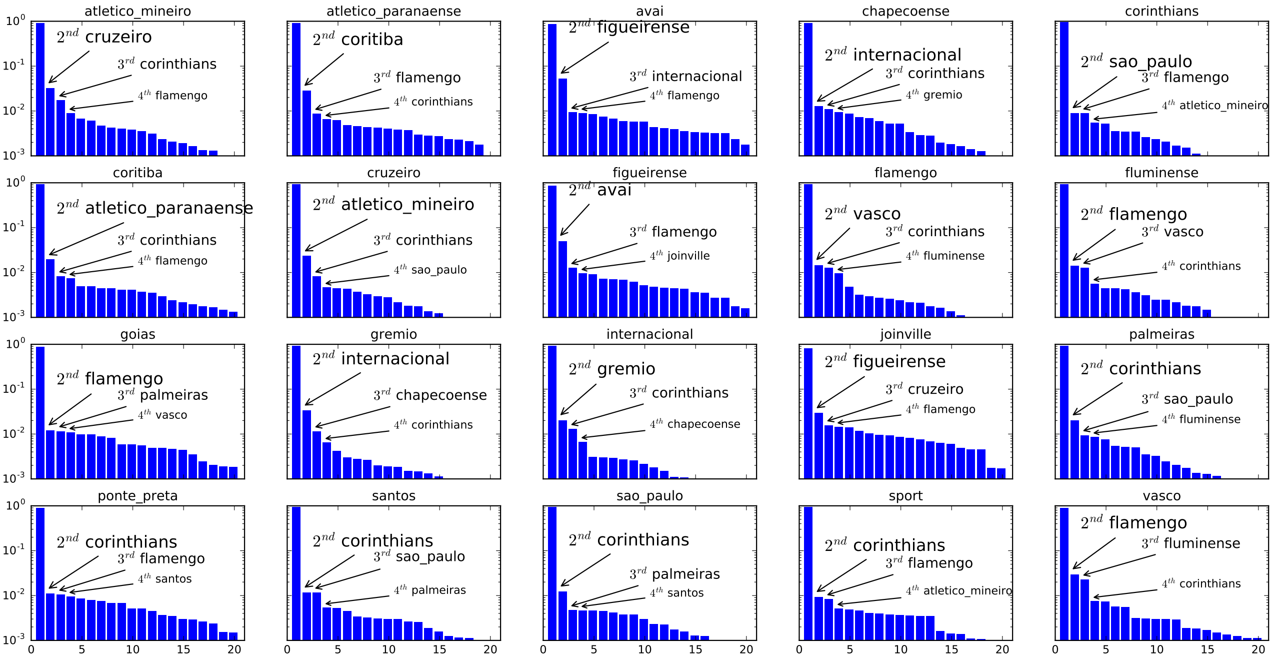

Fig. 3. Clubs ranked distributions ($\mathbf{k}_j^r$) highlighting the top-3 rivals of each club in 2015 Brazilian “Série A”.

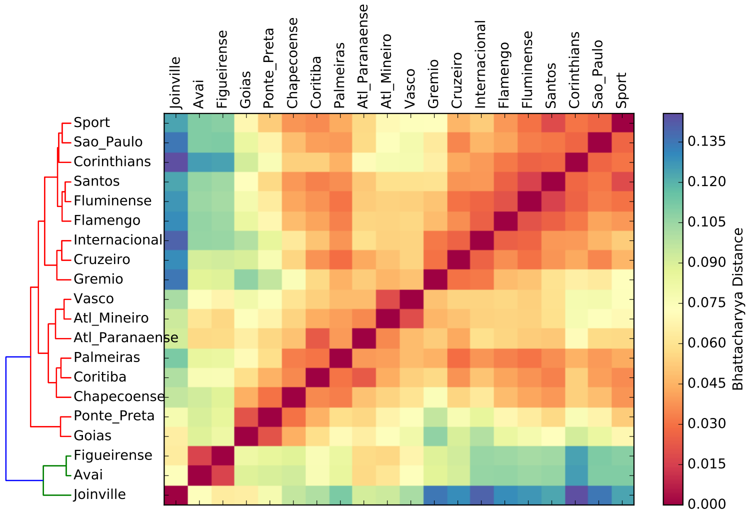

Fig. 4. Heatmaps with hierarchical cluster from 2015 BSA clubs ranked distributions ($\mathbf{k}_j^r$) distributions.

is how competitive are the clubs, i.e how their supporters perceive rivalries among them.

Figures 3 and 5 can aid understanding clusters' rules in figures 4 and 6. For instance, the cluster of clubs with more evenly mentions to all others (i.e. no specific rivalry) has 2 clubs in Brazil – Ponte Preta (11^{th}) and Goias (19^{th}), but 5 in UK – Southampton (7^{th}), Crystal Palace (10^{th}), Stoke (9^{th}), QPR (20^{th}), and Swansea (8^{th}). So, in terms of rivalry, we have 18 and 15 clubs for BSA and EPL, respectively.

On the other hand, the cluster of clubs with highest level of individual rivalries has 3 clubs in Brazil – Figueirense (16^{th}), Avai (17^{th}), and Joinville (20^{th}), and 4 clubs in UK – Man. City (2^{nd}), Arsenal (3^{rd}), Man. United (4^{th}), and Chelsea (1^{st}). Despite clusters being formed based on levels of rivalry, coincidentally, clubs in these clusters are also each other top-rivals. Thus, in terms of rivalry, they form a separated group.

Therefore, at least by the supporters perception, both leagues have two groups of mutual exclusive competitors. These subgroups contain 15 and 3 clubs in Brazil, and 11 and 4 clubs in UK. However, the smallest group in EPL are in the top positions while in BSA they are in the bottom of table. This could explain the discrepancy in number of different champions: 5–EPL and 12–BSA, from 1992-2015.

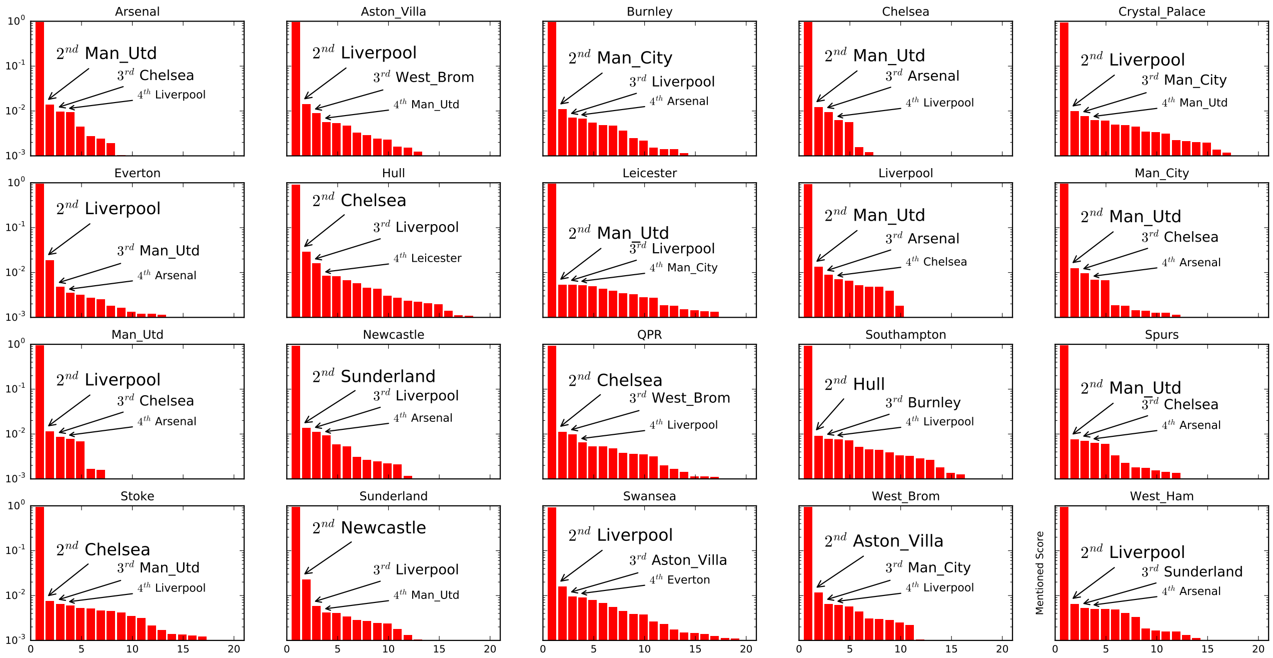

Fig. 5. Clubs ranked distributions ($\mathbf{k}_j^r$) highlighting the top-3 rivals of each club in 14/15 English Premier League.

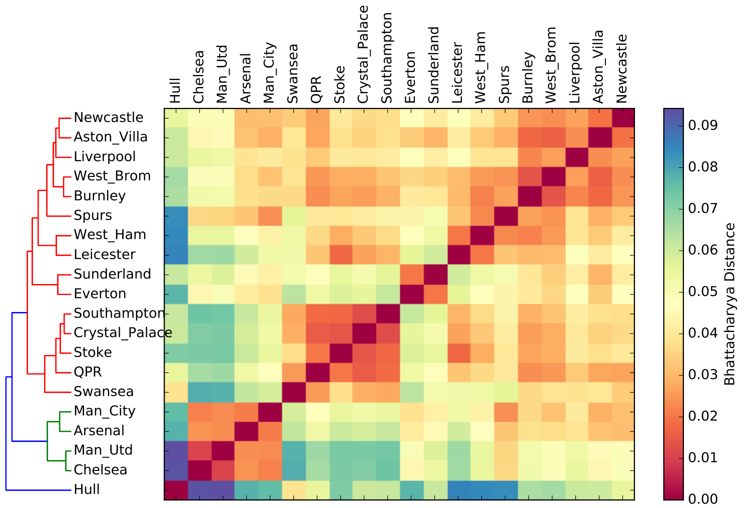

Fig. 6. Heatmaps with hierarchical cluster from 14/15 EPL clubs ranked distributions ($\mathbf{k}_j^r$) distributions.

V. Discussion and Conclusion

In this work we defined an approach to characterize Twitter users as football supporters. Based on the number of times users mention clubs, we can represent these fans by feature vectors, where each component gives the mentioned score for a particular club. The aggregate characterization of several supporters (their vectors) leads to clubs characterization.

We proposed two measurements for clubs: popularity and opposition indexes. The former was defined by the total amount of attention draw to a club. In order to define the later (opposition index), we needed to label users according to their support, i.e. we assumed a user as being fan of the club to which he/she devotes more attention. Thus, we formalized the opposition index as being the total of attention received from non-fans. Finally, we characterized clubs by aggregating their supporters.

We applied the proposed ideas in two different datasets of tweets, representing the 2015 Brazilian “Série A” (BSA) and the 2014/2015 English Premier League (EPL). Together, both datasets contain more than 12.5 million tweets from 1.5 different users.

The results were presented in 3 blocks. First, we found our ranks based on popularity and opposition to be highly correlated to ranks measuring size of supporters ($\hat{r} = .78$) and most-hated clubs ($\hat{r} = .69$). We believe our ranks can be a powerful alternative tool to standard polls, since they require less costs and can be used in more dynamic scenarios.

Second, we ranked the clubs distributions to get supports signatures. Visually, one can understand levels of rivalry and identify main battling clubs. We also noticed a non-symmetric behavior on rivalry, neither for intensity, nor for pairs of clubs.

Last, we used a divergence metric to calculate similarities among clubs support signature and cluster them hierarchically. As these signatures show how supporters behave (intensity distribution of mentions), clustering would reflect this behavior. For instance, we interpreted groups of clubs by growing levels of rivalry. We found a rivalry group in BSA (15) larger than in EPL (4) suggesting a correlation with the number of different champions in these leagues.

The results indicate that the proposed approach can also characterize similar entities in other sports and in other domains since it is a reliable and robust data-drive analysis tool. Our methodology also has the potential to discover meaningful relationships (e.g. support and rivalry in sports domain) unknown *a priori*.

Performing a constrained analysis at a lower granularity (time-span or geography) can aid, for example: (i) decision makers on security issues such as to increase the police contingent for specific matches; (ii) sponsors on targeted marketing, for instance, pointing to regions with expandable supporters or warning at ones with ascending market competitors; and (iii) club's managers to track fans engagement in order to increase attendance at stadiums. These possibilities are tangible and highly useful for practical use and can be obtained in a swift and real-time manner.

VI. Future Works

There are still many questions to be addressed by this work. We list bellow some of them:

- Hashtags are very popular in Twitter. Can we improve or add unbiased ways to identify mentions to clubs?
- Can we have an automatic way to define a user's main club rather than choosing the greater component in the user vector? Maybe with the use of sentiment analysis.
- Can we define other approaches to characterize clubs? What are the consequences on clustering if we use different approaches for characterization?
- Can we get as good results as we got for football if we focus other sports?

In addition, a temporal analysis can show how stable are the supporters, or how they correlate with wins and losses. This could also reveal behavioral differences during the season (beginning, mid, or final) as matches become more decisive.

Yet, the inclusion of sentiment analysis could give a deeper characterization (other behaviors) such as excitement after victories, confidence after long winning strikes, deception when loosing a derby, or even indifference breaks.

Last, we hope that characterization approaches, such as the proposed, can lead to a better understanding of the role of football (and other sports) to society and how it acts as a proxy for certain aspects of the society such as violence, social-economic differences, etc.

Acknowledgments

Diogo Pacheco and Diego Pinheiro supported by CAPES–Brazil, grants 0544/13-2 and 0624/14-4. Ronaldo Menezes and Eraldo Ribeiro partially supported by NSF 1152306.

References

[1] R. Guilianotti, *Football, violence and social identity*. Routledge, 2013.

[2] K. Gwinner and S. R. Swanson, "A model of fan identification: Antecedents and sponsorship outcomes," *Journal of services marketing*, vol. 17, no. 3, pp. 275–294, 2003.

[3] J. Oppenhuisen and L. van Zoonen, "Supporters or customers? fandom, marketing and the political economy of dutch football," *Soccer and Society*, vol. 7, no. 1, pp. 62–75, 2006.

[4] B. Lowe and D. Laffey, "Is Twitter for the Birds? Using Twitter to Enhance Student Learning in a Marketing Course," *Journal of Marketing Education*, vol. 33, no. 2, pp. 183–192, 5 2011.

[5] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185, 2010.

[6] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "Analyzing twitter for social TV: Sentiment extraction for sports," *CEUR Workshop Proceedings*, vol. 720, 2011.

[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?" in *the 19th international conference on World wide web*. New York, New York, USA: ACM Press, 4 2010, pp. 591–600.

[8] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. V. Hentenryck, J. Fowler, and M. Cebrian, "Rapid assessment of disaster damage using social media activity," *Science Advances*, vol. 2, no. March, pp. 1–12, 3 2016.

[9] M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto, "Do altmetrics work? Twitter and ten other social web services," *PloS one*, vol. 8, no. 5, p. e64841, 1 2013.

[10] A. Gruzd and M. Goertzen, "Wired academia: Why social science scholars are using social media," in *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE, 1 2013, pp. 3332–3341.

[11] S. Kumar, G. Barbier, M. A. Ali Abbasi, and H. Liu, "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief," *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 661–662, 2011.

[12] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks." *Science (New York, N.Y.)*, vol. 301, no. 5634, pp. 827–9, 8 2003.

[13] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[14] J. Price, N. Farrington, and L. Hall, "Changing the game? The impact of Twitter on relationships between football clubs, supporters and the sports media." *Soccer & Society*, vol. 14, no. 4015, pp. 446–461, 7 2013.

[15] R. Coche, "Promoting women's soccer through social media: how the US federation used Twitter for the 2011 World Cup," *Soccer & Society*, vol. 17, no. 1, pp. 37–41, 5 2014.

[16] S. Zhao, L. Zhong, J. Wickramasuriya, V. Vasudevan, R. LiKamWa, and A. Rahmati, "SportSense: Real-Time Detection of NFL Game Events from Twitter," *arXiv.org*, vol. 1205, p. 3212, 5 2012.

[17] D. F. Pacheco, F. B. Lima-neto, L. G. Moyano, and R. Menezes, "Football Conversations: What Twitter Reveals about the 2014 World Cup," in *Brazilian Workshop on Social Network Analysis and Mining (CSBC 2015 - BraSNAM)*, Recife, 2015.

[18] A. Lehman, N. O'Rourke, L. Hatcher, and E. Stepanski, *JMP for Basic Univariate and Multivariate Statistics: Methods for Researchers and Social Scientists, Second Edition*. SAS Institute, 2013.

[19] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 2 1967.